The Theoretical Status of Ontologies in Natural Language Processing

John A. Bateman[1]
Projekt KOMET and Penman Project
GMD/IPSI and USC/ISI
e-mail: bateman@gmd.de

March 5, 1992

Note: *This paper appears in the Proceedings of the workshop on 'Text Representation and Domain Modelling – Ideas from Linguistics and AI', held at the Technical University Berlin, October 9th - 11th, 1991. KIT Report 97, edited by Susanne Preuß and Birte Schmitz.*

Cmp-lg Paper No: cmp-lg/9704010

[1]Most of the background for this paper is drawn from experiences with the development of the Penman *Upper Model*: an ontology for supporting natural language generation. The Upper Model has been under development since 1985, and many have and continue to contribute to it. The ideas I report on here would not have been possible without that development. Those responsible for the present form of the Upper Model include: William Mann, Christian Matthiessen, Robert Kasper, Richard Whitney, Johanna Moore, Eduard Hovy, Yigal Arens, and myself.

1 Introduction

The development of natural language processing (henceforth, NLP) systems has reached the stage where concentrated efforts are necessary in the area of representing more 'abstract', more 'knowledge'-related bodies of information. It has been accepted that without substantial bodies of background information concerning commonsense, everyday knowledge about the world or detailed information concerning particular domains of application, it will not be possible to construct systems that can support the use of natural language. Systems need to represent concrete details of the 'worlds' that their texts describe: for example, the resolution of anaphors, the induction of text coherence by recognizing regularities present in the world and not in the text, the recognition of plans by knowing what kinds of plans make sense for speakers and hearers in real situations, etc. all require world modelling to various depths.

This need creates two interrelated problem areas. The first problem is how knowledge of the world — be it general, commonsense knowledge or specialized knowledge concerning some particular domain — is to be represented. The second problem is how such organizations of knowledge are to be related to linguistic system levels of organization such as grammar and lexis. For both problem areas the concept of **ontologies** for NLP has been suggested to be of potential value. Very generally, an ontology offers a 'conceptual' framework for the representation of information — a framework that is sufficiently general, but also sufficiently detailed, to provide a rich supportive scaffolding for the construction of models of the world. The design of such ontologies constitutes an area of concern that is coming to be known as *ontological engineering* (e.g., [Nirenburg and Raskin, 1987, Lenat and Guha, 1988, Simmons, 1991]. As we shall see below, most systems that deal currently with NLP already adopt some kind of ontology for their more abstract levels of information. However, theoretical principles for the design and development of ontologies meeting the goals of generality and detail remain weak. This is due not only to a lack of theoretical accounts at these more rarified abstract levels of information, but also to the co-existence of a range of, sometimes poorly differentiated, functions such bodies of information are expected to fulfill.

The following list gives an idea of the range of functions adopted in NLP. Ontologies are often expected to fulfill at least one (and often more) of:

- organizing 'world knowledge',
- organizing the world itself,
- organizing 'meaning' or 'semantics' of natural language expressions,
- providing an interface between system external components, domain models, etc. and NLP linguistic components,
- ensuring expressability of input expressions,
- offering an interlingua for machine translation,
- supporting the construction of 'conceptual dictionaries'.

Moreover, an ontology is seen as a very *general* organizational device: i.e., one that provides a classification system for whatever area of application the ontology is applied to. The organizational resource offered by an ontology has to be *re-usable*. But it is an open issue as to what extent the kinds of organization listed here overlap. It cannot be taken for granted that they all refer to the same level of abstract description. It can also not be taken for granted that there is unity concerning the *tasks* that are involved in such descriptions. This can be seen in the following statement from Hobbs.

> 'Semantics is the attempted specification of the relation between language and the world. However, this requires a theory of the world. There is a spectrum of choices one can make in this regard. At one end of the spectrum — let's say the right end — one can adopt the "correct" theory of the world, the one given by quantum mechanics and the other sciences. If one does this, semantics becomes impossible because it is no less than all of science... There's too much of a mismatch between the way we view the world and the way the world really is. At the left end, one can assume a theory of the world

> that is isomorphic to the way we talk about it. ... Most activity in semantics today is slightly to the right of the extreme left end of this spectrum. ... it fails to move far enough away from language to represent significant progress towards the right end of the spectrum.' [Hobbs, 1985, p68]

It probably does not make sense, therefore, to talk of a generalized classification system without first fixing more precisely the nature of its intended function. A further problem is that the first of the desired functions above, organizing world knowledge, is often taken to be *definitional* for an ontology.[2] However, the world — i.e., psychological, logical, or philosophical views of the world — has not proved to be very constraining as to what knowledge organizations it requires. 'Ontologies' built on the basis of such constraints are, as we shall see below, underconstrained and there has accordingly been no achievement of the large scale resources necessary for re-use across NLP systems.

The main purpose of this paper is to add a further round of discussion to that concerning the design and construction of ontologies for NLP. The paper is explicitly explorative, building on experience in the definition and use of such ontologies for text generation. The paper is intended to stimulate discussion, rather than present solutions — although I do conclude with suggestions for certain lines of theoretically motivated methodological development for future ontologies. The basic path taken in the paper will be to differentiate among the distinct functions that ontologies may serve in order to be better able to set out principles and constraints for the design of abstract levels of knowledge organization that can serve as ontologies appropriate for NLP. Seen in more detail, the paper is organized as follows.

First, I discuss the role of language as a possible motivating force for designing and populating ontologies. Second, I introduce several of the most extensive ontologies that are currently to be found in NLP systems, characterizing their precise function and motivation within their respective systems. Third, I relate the distinct types of ontology discovered to possible general linguistic theories that would support them. It is my contention that many principles of organization follow directly from the position of suggested bodies of information in the linguistic system as a whole and that recognizing this allows efforts in the definition and construction of such bodies of organization to be directed more appropriately than has hitherto been the case. For any ontology that is proposed, therefore, it should be possible to relate its properties back to a motivating linguistic theory. I argue that the evidence that we now have from the more extensive attempts at ontology construction suggests strongly that a richly *stratified* model of the linguistic system is required in order to achieve the degree of constraint that we need for attacking large-scale, re-usable ontology construction. Fourth, I show how the ontology of the Penman text generation system — that has been developed largely as an instantiation of the highly stratified theory of systemic-functional linguistics — already answers many of the criticisms that have been raised against other ontologies. I argue that although these criticisms are often based on largely *post hoc*, methodological grounds, the vast majority of them also follow directly from the properties of the linguistic system and so could (and arguably *should*) have been made prior to attempting ontology construction. This can be seen in the properties of the Penman ontology, whose very design avoids significant criticisms levelled elsewhere. Finally, I suggest how ontology design could be improved yet further by taking into consideration more input from linguistic theory. The Penman ontology, for example, is only a partial instantiation of the theoretical principles underlying it and it is possible to show that problems enter into the account precisely where the ontology falls short of the theoretical specification.

In general, then, this paper is intended not only to improve our understanding of what kinds of bodies of information can stand as ontologies of various kinds and how such bodies of information relate to other resources in the computational representation of the linguistic system, but also to make the point that appropriate views of the rich dimensions of organization exhibited by the linguistic system can go a long way to improving our initial design specifications for NLP systems. They

[2] Or the second may be claimed to be the real task — however, as Hobbs points out, this actually comes closer to the first position.

should, therefore, always be considered very early on in system construction and computational theory development.

2 The role of language in ontology justification

As mentioned above, the move to consider NLP systems that require information over and beyond that attributable to surface syntax has raised two problems: how to organize that information and how to relate that information with the less abstract levels of the linguistic system. The first problem is typically considered in more detail in approaches where the operation of a system *in some specified domain* is the central goal; the second usually arises in systems which attempt to model the linguistic system itself, focusing less closely on the embedding in any particular specified domain of application.

One common source for knowledge construction and representation that is found in approaches to the first problem is earlier work in artificial intelligence (AI). Even early AI reasoning programs needed to represent the state of the world in which the programs were to operate. This has given rise to the areas of **domain modelling** and **commonsense reasoning** which are responsible for representing concrete details of aspects of the world. The enterprise of world modelling clearly has many similarities with the requirements of sophisticated NLP systems and there has naturally been an influx of techniques and attitudes concerning ontology design from the AI context.

This has proved most successful in the cross-over of techniques of knowledge representation in AI to techniques for representing linguistic information. The similarity between structured inheritance knowledge representation languages such as KL-ONE [Brachman and Schmolze, 1985] and its descendents and current typed feature logics (e.g., [Smolka, 1989, Smolka and Aït-Kaci, 1989, Nebel *et al.*, 1991]) is an active area of research. A basic model for the *representation* of ontologies can now assume minimally that a subsumption lattice over sorts is defined, probably with some mechanism corresponding to the structured inheritance of role information associated with the sorts, and possibly additional axioms, or particular inferences, licenced by specified combinations of sorts. This will be the representational basis for ontologies of all kinds that I will assume throughout this paper.

In contrast to this concensus, attempts to decide exactly *which* sorts make sense for an ontology based on AI 'knowledge engineering' principles have been less successful. Although the effort-intensive nature of domain modelling naturally calls for consideration of the re-usability of components of the knowledge represented across distinct domains, the ability of AI-centered approaches to come up with such general organizations has been limited. Some of the earliest work in this area was that on 'naive physics' (e.g., [Hayes, 1979, Hayes, 1985]): here the aim was to capture the underlying 'general knowledge' that people have about physical objects and substances in the world; similar investigations are reported in, for example, [Hobbs and Moore, 1985, Hobbs *et al.*, 1987], and there are naturally also connections to be drawn with other work in semantic and 'conceptual' representation in AI, e.g., [Schank and Abelson, 1977]. Further good examples of systems that require detailed real 'knowledge' in particular domains are expert systems; here also there is still little shareability across domain models. The detailed organization of such systems' knowledge is typically unique to particular application domains and shows relatively little cross-domain re-usability.

We can in part explain this by considering the relative importance assigned to the distinct functions that such domain models in AI are to fulfill. For example, when constructing a knowledge source whose primary function is to support the particular inferences that a given system needs to draw, it is logical that the organization of that knowledge be tailored with this goal in mind. This usually leads, however, to nongeneralizeable representational requirements because the inferences that distinct systems are to draw have not been related. The relatively small scale of most of this work to date has furthermore limited the effectiveness and urgency of investigations into re-usability: the cost of constructing domain models from scratch has not been

prohibitively high. This cost-equation quickly changes once more realistically sized bodies of information are considered. It quickly becomes much more important that detailed organizations of general knowledge applicable to many domains are available so as to reduce the work involved when moving to new domains.

The most extensive attempt to create a general scaffolding for representing general, background knowledge of the world based on AI techniques is the CYC project [Lenat and Guha, 1988]. The size of this project (initial projections were for a base level of 10,000,000 entries) of necessity forces a sharp awareness of the need to have an organization for knowledge that is detailed and general enough to provide sufficient scaffolding for supporting large-scale bodies of information in accessible and usable ways. Without clear principles both for the organization of such knowledge and for the selection of the information to be represented, the result would be disastrous: poorly organized knowledge will be inadequate both theoretically, in that it fails to capture significant generalizations, and practically, in that it fails to be usable as a resource. The procedure followed in CYC is to divide up types of entities into categories that appear to behave differently, i.e., concepts are classified according to the kinds of inferences that they allow to be drawn about themselves. Problematic here, therefore, is precisely which kinds of inferences are to be taken as definitional. This does not appear to have been made explicit and so the procedure does not provide a particularly sound methodology. The resulting *domain-independent*, and hence re-usable, portion of the CYC ontology is accordingly not very deep, somewhat tangled, and supports limited inferences. It then becomes increasingly necessary to raise questions concerning the consistency of distinct areas of knowledge represented and, consequently, how one can use that knowledge.

It needs to be recognized that it is essential to define the purpose for which a body of information is to be used in order to define appropriate organizations for that information. As long as the purposes are unclear, or too varied, consistent organizations will be difficult to achieve. The statement that a general ontology of real-world knowledge should simply 'represent' that knowledge is underspecified. It does not provide sufficient guidance for finding useful organizations for that knowledge. Given that we need a general organization and that that organization will be determined by purpose, we clearly need a very general (but still formally specifiable) task that requires particular inferences to be performed. If it were possible to find such a task, then it would be possible to use it as a guiding methodology for constructing general organizations of knowledge. Precisely one such general task is, of course, the expression of knowledge in natural language: whatever the knowledge that is represented, i.e., whatever domain and however general/specific, it should be possible to express that knowledge linguistically.[3] One additional set of constraints that one can apply in the construction of organizations of knowledge that attempt maximal applicability across domains is then that offered by language.

This must be specified further. For example, the acceptance of 'ways of talking' about categories as evidence for the existence of those categories in an ontology is a very old strategy (e.g., Aristotle) and is present even in CYC. This method of justification is, however, limited to seeing what one can say and still make sense about a category rather than any more technical analysis of linguistic properties. The precise 'inferences' that are being relied upon to shape the organization are, therefore, still not being given. Thus, there are examples of ontologies that are constructed in NLP systems, where there is a specified relationship between concepts and linguistic expression, but the relationship is sufficiently non-general so as not to provide strong constraints on ontology design.

One such case is the ontology of KBMT projects [Carbonell and Tomita, 1987] such as TRANSLATOR [Nirenburg and Raskin, 1987]. Work of this kind seeks a level of representation that is minimally different across distinct languages. Moreover, the value of organizations of information that are relevant across distinct domains is clearly recognized

[3]This is overstated to the extent that some information/knowledge is often maintained to be inexpressible linguistically — even if this is so, it is still the case that by far the widest and most generally applicable form of expression that we know is language. In any case, whether or not there exists knowledge that is inexpressible linguistically will not affect the final outcome of the discussion below.

and re-usable ontology portions are actively sought. However, although the link to language ensured by the machine translation task increases the likelihood that this can be achieved on a larger scale, the re-usable portions of the ontologies proposed until now remain small. This can in part be attributed to the fact that the appeal made to language as a constraining force on ontology design is undervalued.[4] The 'external-to-language' attitude towards ontological constructs assumed from AI promises to capture abstract models of the world (or of conceptions of the world — a difference that is not criterial at this point) and its organization independent of particular languages. This appears a tempting direction for achieving interlinguality. But then we find 'motivations' such as the following for the categories that are to be adopted within an interlingual ontology:

> 'Russian has no word that corresponds exactly to the English word *afford* (as in *I can't afford X* or I can't afford to Y). In a multilingual processing enviroment, there might be a concept corresponding to a sense of the English word *afford*. A Russian sentence *Ja ne mogu sebe etogo pozvolit' (I can't allow myself this)*, uttered in a context of acquisition ... should involve the concept that represents *afford*. This means that if the units of the representation language are chosen so that they are based on Russian lexis, the meaning of *afford* will be missing. **But this meaning seems sufficiently basic to be included in an ontology.**' [Nirenburg and Levin, 1991] [bold: my emphasis].

It is clear that this kind of argumentation needs to be sharpened considerably; it is also clear that this can only be done when it has been established exactly what function the 'ontology' is to serve. In general, the more detailed the linguistic constraints adopted on ontology design are, the more detailed and *explicitly justifiable* that ontology design becomes.[5] However, the relationship between ontologies and NLP is interestingly reflexive.[6] Ontologies appear necessary for the organization of knowledge appropriately for use by NLP systems, and simultaneously the explicitness of the necessary inferences that constitute an NLP system provide an until now unrivalled source of constraint for deciding on ontology designs.

This connection is described well in the following citation from Ewald Lang:

> '... the structure of language plays a dual role. It is, properly allocated to the parsing and generating components, a constitutive part of the object to be modeled (that is, the system which is to integrate linguistic and non-linguistic knowledge). But at the same time it is also part of the device by means of which this object is accessed, that is, the categorization of lexical items into nouns, verbs, etc., provides an apparently natural grid for establishing corresponding sorts of entities in the ontology, which, by definition, is to represent non-linguistic common sense knowledge. Given this, the risk of confusing linguistic and non-linguistic categories is latent; moreover, it is practically unavoidable as long as we are confined (or confine ourselves) to looking at common sense knowledge through the window of language only, i.e., without a chance to draw on independent evidence from non-linguistic (say, visual or kinasthetic) ways of accessing the structure and contents of common sense knowledge.' [Lang, 1991, p464]

[4] This is also made problematic by the very multilinguality of possible linguistic constraints inherent in machine translation system — without appropriate ways of achieving *linguistic* generalizations across languages (cf., e.g., [Bateman *et al.*, 1991] for discussion), the application of linguistic constraints is very much more difficult.

[5] This was also one result of an extensive study of proposed ontologies reported upon in [Skuce and Monarch, 1990]. Although there has also been at least one example of development that has attempted movement in the opposite direction. The *abstraction structure* of BBN's natural language and understanding project JANUS was redesigned away from a linguistically oriented description in order to find a 'more general ontological style' [Weischedel, 1989, 200] that was not so strongly connected with the linguistic realization of the concepts defined. However, this very move was probably one contributing factor to the less than successful outcome of the subsequent attempt to use the *Longman Dictionary of Contempory English* as the basis for defining a domain-independent taxonomy for JANUS [Reinhardt and Whipple, 1988]. The most significant generalizations that would have helped organize the taxonomy for the purposes of natural language processing had probably already been lost.

[6] Or even circular: as I shall mention below.

Thus, while linguistic patterns are probably the richest source of organizational criteria that are available to ontology design, their use is certainly not unproblematic. Consequences of this can be seen in the fact that although the majority of recent and currently planned natural language processing systems recognize the necessity of some level of abstract 'semantic' organization similar to an ontology that classifies knowledge explicitly according to its possibility for linguistic expression,[7] very few have achieved ontologies of any size and *motivations* for inclusion of particular concepts and distinctions in ontologies remain limited or underspecified. Thus, the decision to use linguistic evidence by itself is still, unless further restricted, underspecified and leaves open a range of positions. These give rise to differing functionalities that the ontologies are to serve, which hence impacts on ontology design. The positions and functionalities need to be characterized more precisely and this I attempt in the following section.

3 Three kinds of ontologies

Although I have concentrated until now on preliminaries to the first problem area mentioned in the introduction — how knowledge of the world is to be represented — the apparent value of applying linguistic constraints to this task renders the second problem area — how that knowledge is related to language — crucial. If the ontology cannot be related to language in an explicit, formalized fashion, then the structures (and functions) of language will be prevented from having a direct constraining influence on what gets represented in the ontology, what not, and how the entire ontology is to be organized.

[7] Including, for example: the Functional Sentence Structure of XTRA: [Allgayer *et al.*, 1989]; [Dahlgren *et al.*, 1989]; [Emele, 1989]; the POLYGLOSS project: [Emele *et al.*, 1990]; certain of the domain and
text structure objects of SPOKESMAN: [Meteer, 1989]; TRANSLATOR: [Nirenburg *et al.*, 1987]; the Semantic Relations of EUROTRA-D: [Steiner *et al.*, 1987]; the JANUS project: [Weischedel, 1989]; and the ontological types of the ACORD project: [Moens *et al.*, 1989]. Moreover, ontology-like organizations of informations have also been found useful for parsing applications by, e.g., [Calder *et al.*, 1989, Chen and Cha, 1988, Hinrichs *et al.*, 1987, Zajac, 1989]. There are no doubt many other places where this kind of construct now appears.

There are at least two theoretically distinct standpoints from which this second problem area has been addressed in NLP systems. One possibility is to assume that real-world domain knowledge is more or less directly linked to grammatical and lexical forms of expression. The organization of the world knowledge ontology should then, ideally, also be supportive of the use of that knowledge for linguistic expression or for interpreting linguistic distinctions: the problem of relating knowledge to language is thus subordinated to the world knowledge ontology design. A second possibility is to assume that the relationship between real-world domain knowledge and grammar and lexis is itself complexly structured. This structuring may lean for its organization towards the world knowledge ontology, in which case this would blend into the first possibility, or towards the grammar and lexicon, or alternatively could rely on its own principles of organization. Each of these variants has been adopted in some system where a concrete ontology has been attempted. This gives rise to three distinct kinds of ontology that can be found in NLP work. An ontology can be

- an abstract semantico-conceptual representation of real-world knowledge that also functions as a semantics for use of grammar and lexis — this type I will term a *mixed* ontology: O_m;
- an abstract organization underlying our use of grammar and lexis that is separate from the conceptual, world knowlege ontology, but which acts as an interface between grammar and lexis and that ontology — this type I will term an *interface* ontology: O_i;
- an abstract organization of real-world knowledge (commonsense or otherwise) that is essentially *non-linguistic* — this type I will term a *conceptual* ontology: O_c.

The relationship involved here, their embedding in general architectures, and the subtypes of interface ontologies mentioned above are depicted graphically in Figure 1.

3.1 Conceptual ontologies

Most of the AI designed ontologies — including

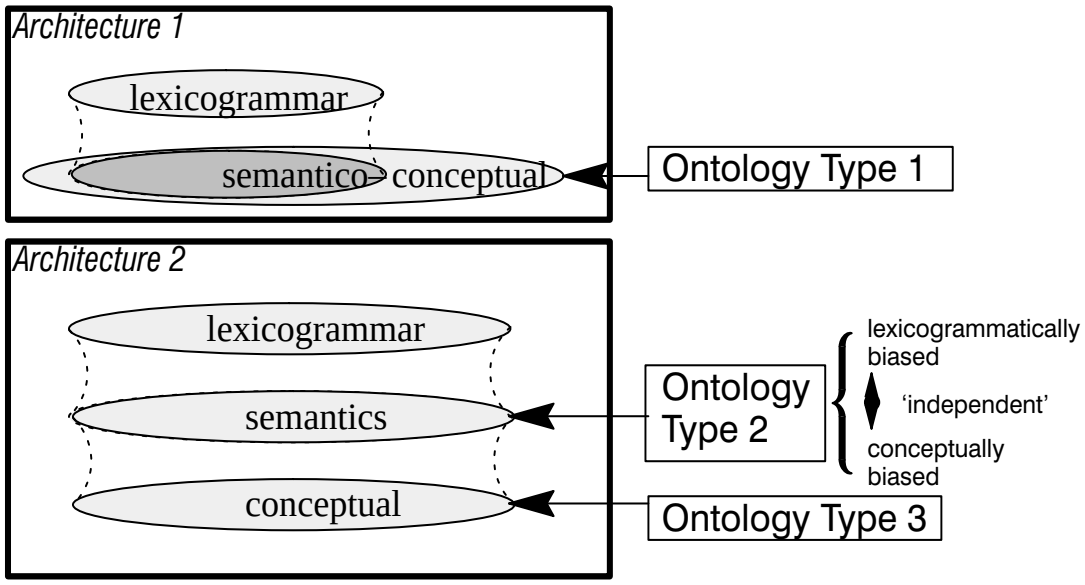

Figure 1: Three kinds of ontology in NLP

those of CYC, TACITUS [Hobbs *et al.*, 1987], JANUS [Weischedel, 1989], 'the naive semantics' of [Dahlgren *et al.*, 1989], and even some aspects of the KBMT ontology, e.g., [Nirenburg and Raskin, 1987, Nirenburg and Levin, 1991] — are attempts to construct ontologies of the third type: pure maximally language independent ontologies reflecting the structure of the world. I have already discussed some of the difficulties of designing such ontologies without building up through an account of language. Psychological research might offer another source of evidence for such ontologies; as would detailed sociological work on the commonsense world. It is, however, unclear whether any such methodology will be able to avoid the relationship to language observed above and so I will now concentrate on ontologies which are at least intended to be related explicitly to language.

3.2 Mixed ontologies

An example of a mixed ontology — i.e., one where there is no extensive treatment of the relation between the world knowledge ontology and grammar and lexis maintained separate to the ontology itself — is the approach taken in the LILOG natural language understanding project [Herzog and Rollinger, 1991]; details of the ontology are given in, for example, [Klose and von Luck, 1991, Pirlein, 1991, Klose *et al.*, 1991], and details of the relation between linguistic forms and conceptual representations are given in, for example, [Gust, 1991, Bosch, 1991]. It may at first glance appear strange to classify LILOG here, since the approach to the relation between linguistic form and world knowledge draws heavily on [Bierwisch, 1982]'s theory of semantics where, to cite [Gust, 1991]'s statement of Bierwisch's position: '... semantic forms and conceptual structures belong to different and strictly discriminated levels'. [Bosch, 1991] also makes it very clear that he holds this distinction to be crucial for making progress in semantics and knowledge representation. However, when the modelling of the approach is examined, we find that this distinction of levels comes under attack. For example, both semantic forms, which are derivable from the lexicon and from grammatical analysis, and conceptual forms are represented in a single language (the se-

mantic language being a subset of the conceptual language: [Bosch, 1991, p248]) and are freely combinable; moreover [Gust, 1991, p133] maintains that: 'there are continuous variations between semantic forms and conceptual structures.' This gives rise to lexical entries which directly contain categories of an ontology which also contains categories of real-world knowledge. The relation between conceptual knowledge and grammatical and lexical form is thus handled by logically manipulating categories from a single ontology until categories are found that possess links to grammatical or lexical entries — this is precisely the architecture consistent with a mixed ontology as shown in Figure 1.

An illustration of the nonseparation of of 'linguistic' information and 'conceptual' information typical of ontologies of this type can be seen in the following taken from [Bosch, 1991, p251]. In order to find the interpretation in context of the lexeme "school" as it is used, arguably differently, in examples such as:

a. The school made a major donation.
b. The school has a flat roof.

A general 'lexical semantic entry' for the lexeme is retrieved thus:

```
SEM(''school'') = λX [PURPOSE X
W]
where
W =
PROCESSES_OF_LEARNING_AND_TEACHING
```

This is then interpreted by applying a given 'contextualizing function' selected depending on the basis of the semantic interpretation of the predicate in the lexicogrammatical representation. Those for the example sentences would be:

```
a.  λX [INSTITUTION X & SEM X]
b.  λX [BUILDING X & SEM X]
```

Combining the semantic entry and the contextualizing function gives the required 'conceptual' concept that is the referent of the lexeme in context — i.e., that "school" is interpreted as either an institution or a building. All of the undefined predicates found in these logical expressions (e.g., INSTITUTION, PURPOSE, etc.) are sorts defined in the ontology. A direct link is therefore constructed from lexicogrammatical information and chunks of information appropriate for the conceptual level of organization. As we shall see in Section 4.2, this direct linking is a common property of NLP systems based on the common notion of 'semantics' and arises out of a view of the linguistic system that collapses together several important distinctions.

3.3 Interface ontologies

The second and third types of ontology — the interface and conceptual types — usually occur, at least theoretically, in the same architecture. Although it is also the case that some systems address themselves to the organization of the interface ontology without specifying how the conceptual ontology will look. This latter position is common for systems that are intended as general purpose NLP systems reusable across different domains and applications. Examples of such systems include both parsers and generators such as the Penman system [Mann and Matthiessen, 1985, Penman Project, 1989] and Mumble-86 [Meteer *et al.*, 1987]. Here the problem of how to organize the interface with external applications, where those applications are not known in advance, has naturally focused attention on organizations of information appropriate for interfacing. The approach to this developed within the Penman project in terms of the **Upper Model** has become more or less typical of how this is achieved — although to what extent this architecture has arisen independently across systems is unclear. The initial formulation of the Upper Model was based on work by M.A.K. Halliday [Halliday, 1982], William Mann and Christian Matthiessen.[8]

The general statement of the interface problem for NLP systems is that machine-internal information needs to be related to strategies for expressing that information in some natu-

[8]The development of the Upper Model ontology, from its inception as the *Upper structure* of the JANUS project of ISI and BBN, up to its inclusion as a standard component of the current Penman text generation system is covered by the following research reports: [Mann, 1985, Mann *et al.*, 1985, Moore and Arens, 1985, Bateman *et al.*, 1990]. The first detailed theoretical precursor to the ontology was set out in 1985 by Halliday and Matthiessen as a general organization for an experiential semantics: this was called the **Bloomington Lattice**. The subsequent development of the Upper Model has deviated somewhat from the purely linguistically motivated work; this will be discussed in more detail below.

ral language. This could be done in a domain-specific way by coding how the application domain requires its information to appear. This is clearly problematic, however: it requires detailed knowledge on the part of the system builder both of how the generator controls its output forms and the kinds of information that the application domain contains. A more general solution to the problem of defining a mapping between knowledge and its linguistic expression is to provide a classification of any particular instances of facts, states of affairs, situations, etc. that occur in terms of a set of general objects and relations of specified types that behave systematically with respect to their possible linguistic realizations. This classification has itself many of the properties of an ontology, e.g., it is a hierachical organization of sorts and roles — although by virtue of its motivation in linguistic realization, it must be seen as a strictly linguistically motivated ontology. Examples here include aspects of [Meteer, 1989]'s *Text Structure Objects* in the SPOKESMAN text generator:

> '[i]t is important to remember that Text Structure objects reflect the semantic type of the *expression* of the information in an object, not some intrinsic type of the object itself.'[Meteer, 1989, p21];

also the ontology of the ACORD system:

> '... the aim of the sort system is not to reflect the characteristics of real world objects and events referred to by linguistic expressions, but rather to systematize the ontological structure evidenced by linguistic expressions' [Moens *et al.*, 1989, p178];

and, of course, the Upper Model of the Penman system that I will describe in more detail below.

The position that such an interface ontology holds between surface details of a language and more abstract knowledge is, however, an uneasy one. As suggested above, it is possible to differentiate among such ontologies according to whether they orientate themselves more towards less abstract or towards more abstract levels of representation. This brings with it two potential problems in ontology design:

- the ontology can be too *shallow*, in that it's categories are a too direct recoding of linguistic distinctions that do not achieve a qualitative increase in abstraction;
- the ontology can be too *deep*, in that it is no longer possible to draw any formally specifiable connection between the constructs posited and the linguistic evidence taken as motivating them.

Both extreme situations occur and both reduce the value of the ontology as an effective interface medium. The former problem will be accompanied by an increased difficulty in linking the ontology to information of particular domains — regardless of whether this information is considered as a separate kind of information or as more specific details of the same kind of information; and the second problem will be accompanied both by an increased difficulty in linking with grammar and lexis and by the problems induced by poorer linguistic constraints mentioned above. The latter situation then often places a heavier reliance on 'internal' or formal constraints on organization (cf., e.g., [Weischedel, 1989, Horacek, 1989] and what [Lang, 1991, p468] terms 'sortal' restictions) which, while important, do not provide sufficient grounds for deducing very much detailed actual *content* by themselves.

3.3.1 Interface ontologies that are not abstract enough

Interface ontologies exhibiting the former problem are very common and so it is worthwhile giving a slightly more detailed example of the problems that arise. One such ontology is that constituted by the semantic relations used within the german component of the EUROTRA project [Steiner, 1987, Steiner *et al.*, 1987,
Steiner and Reuther, 1989]. These relations are a further development of earlier work by Fawcett — particularly his work on *transitivity* in English (e.g., [Fawcett, 1987]). Fawcett proposes a semantically motivated taxonomy of process types, analogously to the approach taken in [Halliday, 1985] but differing in the actual categories adopted. Each process type has some distinctive set of possible participants — the approach thus differs from early accounts of semantic participants, such as Case Grammar [Fillmore, 1968], where the participant relationships were often defined

separately from the processes in which they participate, and further articulates conceptions of 'thematic' relations such as those found in Lexical-functional grammar (cf., [Hale and Keyser, 1986,
Levin, 1987]) and Government and Binding theory [Jackendoff, 1987]. The EUROTRA-D work has made refinements to the proposed taxonomy on the basis of multi-lingual evidence, particularly from German, so as to provide explicit *syntactic* tests for the assignment of processes to each of the various process types. It is then explicitly stated that the resulting process types described are *no longer* primarily semantic since their classification is based exclusively on differentiation by syntactic criteria. Therefore, although this has produced a framework within which processes can be classified according to the given taxonomy with a high degree of inter-coder consistency, which is an important criterion in large distributed projects such as EUROTRA, its effectiveness as a step towards a higher level of abstract information has been restricted. This can be seen in the following example of process classification given in [Steiner and Reuther, 1989]. For the clause

> That she gave no answer means that she agrees with the proposal.

both subject and object are realized by *that*-clauses and the only possible classification according to the syntactic tests is then one of a *mental process* with two *phenomena*. However, semantically the process also has strong elements of a *relation* between the propositions involved. Similar examples in German are the following: the verb *retten ... vor*:

> Daß er gut schwimmen konnte, rettete ihn vor dem Ertrinken.
> *That he could swim well saved him from drowning.*

again resembles a relational process but has to be assigned to mental according to the criteria formulated; the process 'reden' (*to speak, talk*), which would intuitively seem to be some kind of communication verb, cannot enter into constructions of the form:

> * Peter redet: Karl kommt morgen
> *Peter speaks: Karl is coming tomorrow*

and so does not receive a communication verb classification: and the form:

> * Peter redet, daß Karl kommt
> *Peter speaks that Karl is coming*

cannot occur so it may not even receive a mental reading — the only acceptable forms possible, e.g.:

> Peter redet Unsinn
> *Peter speaks nonsense*
> Peter redet mit Paul *Peter speaks with Paul*

require an action classification, just as the corresponding English processes would. These problems provide evidence that the syntactic tests need to be made more subtle or more elaborate in order to be able to reveal semantic distinctions more reliably. In addition, there is no account suggested of how this level of representation can link to more abstract levels of representation such as a conceptual ontology.

A similar case of this probably contributes to some of the difficulties that arise with the use of 'Lexical Semantic Structures' (LSS) and [Jackendoff, 1983]'s 'Lexical Conceptual Structures' (LCS) for translation — the former as described by [Dorr, 1991], the latter by [Nirenburg and Levin, 1991]. Both structures are tightly bound to possible surface forms by formally specified *linking rules* (e.g., [Levin, 1987]). These rules partition the LSS or LCS into classes reflecting the different realizational behaviour of their categories. Although it is also then sometimes possible to assign to these classes particular 'semantic' features this has still not yet been found to be sufficiently abstract to support a motivated construction of the corresponding conceptual ontology — as the example of the motivation for including the concept *afford* for Russian that I cited above shows. The final selection of conceptual ontological sorts in this case then shows similarities both with that described for LILOG: i.e., by applying a mixture of lexical, grammatical, and domain knowledge criteria, and with the pure AI techniques of CYC and others. In the longer term, therefore, similar problems will occur.

As a final example of the problems of lack of abstraction, I will mention some that have arisen in our development and use of the the Penman Upper Model. The Upper Model, for reasons that I will describe below, does succeed in being more abstract than the semantic relations adopted, for example, within

EUROTRA-D. The organization of an Upper Model achieves greater *semantic* coherence, grouping together distinctions that may be used by a variety of distinct grammatical resources in a grammar. For example, the relationships between process and participants may drive the organization of clauses, but they may equally drive the organization of head and modifiers in nominal groups. The nature of the process-participant relationships is not, arguably, altered by their realizational form. Upper Model generalizations might then express the commonality that unites the following area of variation:

A shoots B
B was shot at T
the shooting of B by A
A's shooting of B
B's shooting
the shooting at P
the P shooting
the T shooting
etc.

under a single specification:[9]

process:
shoot(murderer:A,murdered:B,time:T,place:P).

Categories in the Upper Model are then capturing generalizations which are not appropriately expressed within the grammar. A further example drawn from the 1989 version of the Upper Model is the possible grammatical realizations of the concept of *generalized possession*. This concept should be seen as being realized by possible selections from all the grammatical systems to do with 'possession'. Thus the semantics of the following forms all make reference to this single Upper Model concept.

the door's handle
the handle of the door
the handle that the door has
the door handle
the handle is part of the door
etc.

For a more extensive sets of examples of the lexico-grammatical variation that the Upper Model is intended to support, see [Bateman, 1989, Bateman, 1990a].

[9]Although the actual representation used in the Upper Model reifies both predicates and the relations holding between predicates and their arguments; cf. [Mann *et al.*, 1985, Hobbs *et al.*, 1987, Bateman *et al.*, 1990].

This means that the Upper Model does succeed in achieving a sufficiently high degree of abstraction as to be useful as an interface medium. This increase in abstraction also makes the ontology better suited to linking with more abstract levels of information.[10] The Penman system has been successfully interfaced with a number of applications — mostly expert systems, but also text planners — where domain knowledge is represented. It is then an example of an ontology that mediates the relationship between lexico-grammar and world knowledge without losing the necessary formal connection with the grammar and lexis. Moreover, it moves beyond problems such as that recognized for the EUROTRA-D classification of process types that[11]

> 'The classification system proposed by EUROTRA-D proceeds in a strictly syntactic way. ... From the standpoint of generation this solution is problematic: it would be preferable to have a semantic classification that generalizes across such surface syntactic subtleties." [Heid *et al.*, 1988, p158]

To the extent that it is successful, this is precisely what the Upper Model provides. It achieves this by being based very closely not only on a particular, specified grammar — no concepts are admitted into the Upper Model, for example, *unless they have a direct and specifiable consequence for the operation of the grammar* — but also on a grammar which is *itself* already more abstract than a constituency grammar. I shall describe this in more detail below.

The Upper Model thus stands as a significant step forward in dealing with the problem of interfacing with a general NLP system. The Upper Model decomposes the mapping problem inherent in relating domain knowledge with its possibilities for linguistic expression by establishing a level of *linguistically motivated* knowledge organization specifically constructed as a reponse to the task of constraining linguistic realizations. While it may

[10]So much so that it has sometimes been our experience that the domain model of some application domains has been altered in the light of the consistent organization that the Upper Model brings to bear.

[11]This problem arose while attempting to interface the level of input specification for an existing generator of German (SEMSYN, cf.: [Rösner, 1988]) with the EUROTRA-D semantic relations.

not be reasonable to insist that *application domains* organize their knowledge in terms that respect linguistic realizations — as this may not provide suitable organizations for, e.g., domain-internal reasoning — we have found that it *is* reasonable, indeed essential, that domain knowledge be so organized if it is also to support expression in natural language relying on general natural language processing capabilities.

The general types constructed within the Upper Model necessarily respect generalizations concerning how distinct semantic types can be realized. We then achieve the necessary link between particular domain knowledge and the Upper model by having an application *classify* its knowledge organization in terms of the general semantic categories that the Upper Model provides. This should not require any expertise in grammar or in the mapping between Upper Model and grammar. An application needs only to concern itself with the 'meaning' of its own knowledge, and not with fine details of linguistic form. This classification functions solely as an interface between domain knowledge and Upper Model; it does not interfere with domain-internal organization. The text generation system is then responsible for realizing the semantic types of the level of meaning with appropriate grammatical forms.[12] Further, when this classification has been established for a given application, application concepts can be used freely in input specifications since their possibilities for linguistic realization are then known. Interfacing with such a system is thus radically simplified on two counts:

- much of the information specific to language processing is factored out of the input specifications required and into the relationship between Upper Model and linguistic resources;
- the need for domain-specific linguistic processing rules is greatly reduced since the Upper Model provides a domain-independent, general and reusable conceptual organization that may be used to classify all domain-specific knowledge when linguistic processing is to be performed.

An example of the simplification that use of the Upper Model offers for a text generation system interface language can be seen by contrasting the input specification required for generators that work with realization classes that are less abstract than those of the Upper Model — such as, e.g., MUMBLE-86 [Meteer *et al.*, 1987], or unification-based frameworks,
such as [McKeown and Paris, 1987] and the Lexical Functional Grammar (LFG) approach of [Momma and Dörre, 1987] — with the input required for Penman.[13] Figure 2 shows corresponding inputs for the generation of the simple clause: *Fluffy is chasing little mice*. The appropriate classification of domain knowledge concepts such as *chase, dog, mouse*, and *little* in terms of the general semantic types of the Upper Model (in this case, *directed-action, object, object,* and *size* respectively — cf. [Bateman *et al.*, 1990]) automatically provides information about syntactic realization that needs to be explicitly stated in the MUMBLE-86 input (e.g., `S-V-O_two-explicit-args, np-common-noun, restrictive-modifier, adjective`). Thus, for example, the classification of a concept *mouse* as an *object* in the Upper Model is sufficient for the grammar to consider a realization such as, in MUMBLE-86 terms, a `general-np` with a particular `np-common-noun` and `accessories` of `gender neuter`. Similarly, the classification of *chase* as a *directed-action* opens up linguistic realization possibilities including clauses with a certain class of transitive verbs and characteristic possibilities for participants, corresponding nominalizations, etc. Such low-level syntactic information is redundant for the PENMAN input.[14] Similar, illustrative in-

[12] This is handled in the PENMAN system by the grammar's *inquiry semantics*, which has been described and illustrated extensively elsewhere (cf., [Penman Project, 1989]) and see Section 4.2 below.

[13] Note that this is not intended to single out these approaches at all, the problem is quite general and occurs whenever there is no ontology available for organizing information at a more abstract level than that imposed by the grammar. Further, as already noted, most current NLP developments are moving in a direction analogous to that taken in our work on the Upper Model.

[14] Moreover, when additional information is required, that information is supplied in *semantic* terms rather than in terms of morphosyntactic labeling such as `:number plural` — in this case this is represented

puts forms can easily be imagined for other types of syntactically oriented grammar and lexis components.

The further *domain-independence* of the Upper Model is shown in the following example of text generation control. Consider two rather different domains: a navy database of ships and an expert system for digital circuit diagnosis.[15] The navy data base contains information concerning ships, submarines, ports, geographical regions, etc. and the kinds of activities that ships, submarines, etc. can take part in. The digital circuit diagnosis expert system contains information about subcomponents of digital circuits, the kinds of connections between those subcomponents, their possible functions, etc. A typical sentence from each domain might be:

> *circuit domain:* The faulty system is connected to the input.
> *navy domain:* The ship which was inoperative is sailing to Sasebo.

The input specifications for both of these sentences are shown in Figure 3. These specifications freely intermix Upper Model roles and concepts (e.g., *domain, range, property-ascription*) and the respective domain roles and concepts (e.g., *system, faulty, input, destination, sail, ship, inoperative*). Both forms are rendered interpretable by the subordination of the domain concepts to the single generalized hierarchy of the Upper Model. This is illustrated graphically in Figure 4. Here we see the single hierarchy of the Upper Model being used to subordinate concepts from the two domains. The domain concept `system`, for example, is subordinated to the Upper Model concept *object*, domain concept `inoperative` to Upper Model concept *quality*, etc. By virtue of these subordinations, the grammar and semantics of the generator can interpret the input specifications in order to produce appropriate linguistic realizations: the Upper Model concept *object* licenses a particular set of realizations, as do the concepts *quality*, *material-process*, etc.[16]

Despite the progress that has been made with the Upper Model as a potential interface ontology, it is still the case that the mappings between grammatical forms and the categories of the Upper Model ontology are not yet rich enough to ensure entirely appropriate semantic classifications — entirely anologously to the case with the explicitly syntactically oriented categories of the EUROTRA-D semantic relations. In an attempt to make the definitions of the Upper Model concepts more accessible to users of the Penman system, these definitions have been pushed towards an intepretation of the Upper Model as *predominantly* a hierarchy of generalizations about possible linguistic realizations in English. This approach permits a very straightforward control of the grammar but compromises some of the semantic integrity. Some simple examples of this may be seen in the following.

In the then current version of the grammar, the following clause, which is an example that arose during development of Johanna Moore's *Program Enhancer Advisor* (PEA) system [Moore, 1989]:[17]

X is defined as Y

had to be constructed from a process *define* and an adjunct of 'role-playing' to produce the prepositional phrase *as Y*. This contrasts with a more semantically oriented discrimination of process types which could take, perhaps, a process of 'defining' with three necessary participants, a *definer*, a *defined*, and a *definition*, and state how these are realized directly. In the realization class view as we have it now, the process of defining has to be explicitly decomposed semantically at the level of the Upper Model into a process and a relationship of role-playing. This is not intuitively obvious: indeed, a user has to know *how the grammar generates as-prepositional phrases* in order to arrive at the 'correct' Upper Model classification in order to be able to generate the clause. This is dangerously close to the amount of low-level syntactic detail that needs to be provided for a Functional Unification Grammar or Mumble-86.

in inquiry semantics by the inquiry response pairs {:multiplicity-q multiple} and {:singularity-q nonsingular}. This is also the case for 'tense' but I have abbreviated the semantic specification here. For descriptions of all these distinctions in detail, see the PENMAN documentation [Penman Project, 1989].

[15] These are, in fact, two domains with which we have had experience generating texts using the Upper Model.

[16] For further discussion of this simplification in the semantic input specification for the sentence generator, see [Bateman, 1990b].

[17] All the PEA examples were provided in work by Johanna Moore and Richard Whitney.

```
(general-clause
  :head (CHASES/S-V-O_two-explicit-args
          (general-np
            :head (np-proper-name "Fluffy")
            :accessories (:number singular
                          :gender masculine
                          :person third
                          :determiner-policy no-determiner))
          (general-np
            :head (np-common-noun "mouse")
            :accessories (:number plural
                          :gender neuter
                          :person third
                          :determiner-policy initially indefinite)
            :further-specifications
              ((:attachment-function restrictive-modifier
                :specification (predication-to-be *self*
                                  (adjective "little"))) )) )
  :accessories (:tense-modal present :progressive
                :unmarked) )
```

Input to MUMBLE-86 for the clause: *Fluffy is chasing little mice*
from: [Meteer *et al.*, 1987]

```
(e / chase
   :actor (e / dog :name Fluffy)
   :actee (m / mouse
               :size-ascription (s / little)
               :multiplicity-q multiple :singularity-q nonsingular)
   :tense present-progressive)
```

Corresponding input to PENMAN

Figure 2: Comparison of input requirements for MUMBLE-86 and PENMAN

```
(v1 / connects
    :domain (v2 / system
                :relations (v3 / property-ascription
                               :domain v2
                               :range (v4 / faulty)))
    :range (v5 / input)
    :tense present)
```

Input for digital circuit example sentence:
The faulty system is connected to the input

```
(v1 / sail
    :actor (v2 / ship
               :relations (v3 / property-ascription
                              :domain v2
                              :range (v4 / inoperative)
                              :tense past)
    :destination (sasebo / port)
    :tense present-progressive)
```

Input for navy example sentence:
The ship which was inoperative is sailing to Sasebo

Figure 3: Input specifications from navy and digital circuit domains

This is not an isolated case. Other problematic assignments in the PEA domain include:

- The process *call*, as in "The boy is called John". Presently *call* is classified as a *dispositive-material-action* from UM-89, *boy* becomes the *actee*, and the name, 'John', becomes a *recipient*. No *actor* is specified and so a passive construction appears (due to a then current shortcut defined for the textual reasoning that the grammar initiates for selection of active-passive clauses).

- The process *generalize to*, as in "The result can be generalized to other cases". Here *generalize* is again a straightforward *nondirected-action* and *to other cases* is specified as a *destination* spatio-temporal circumstance in UM-89 in order to generate the preposition.

In all of these cases, the role assignments are only being used in order to achieve the required syntactic pattern given by the particular state of the grammar of the Penman system: the Nigel grammar of English. In the first example, the model for the clause being used is that of *give* since this class of verbs is bitransitive; in the second, the technique adopted is as with the case of *define as* above, where a circumstantial role is selected purely in order to guarantee the desired preposition. Although in these cases it is reasonably clear that both grammar and Upper Model would need to be extended to include the desired process types, in general the *theoretical* status of using arbitrary assignments of concepts to the Upper Model and selections of roles to be expressed has not been made sufficiently clear. This technique is (or rather *should*) only be employed when it is not possible to extend the grammar and ontology appropriately: only when the grammar has to be taken as 'fixed', e.g., because it is being applied by a user that does not have access to the internal organization of the grammar, is this kind of strategy defensible. As a general technique, the strategy has to be strongly rejected on theoretical grounds. However, note that without a commitment to *semantic* coherence, there is little reason not to use the Upper Model in this way; we have already seen the similar situation in the use of the EUROTRA-D system of semantic relations where commitment

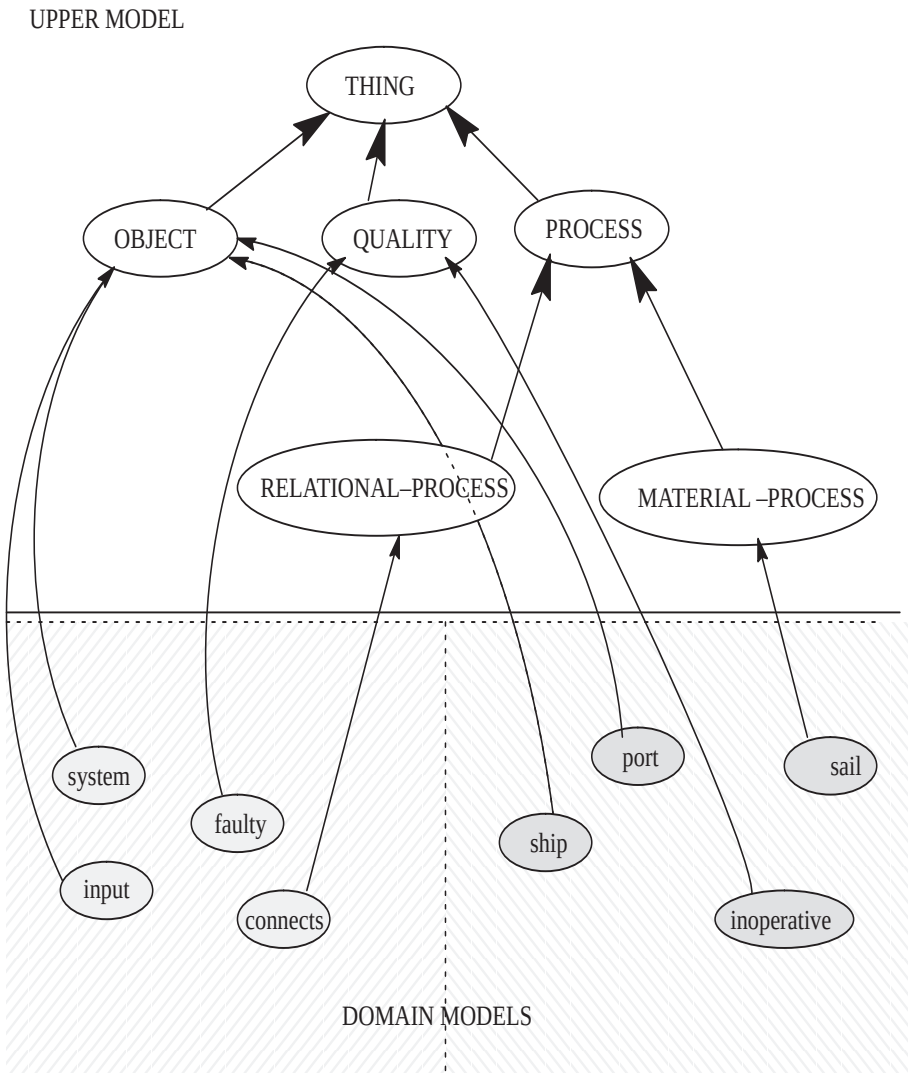

Figure 4: Upper Model organization reuse with differing domains

to semantic coherence has been explicitly rejected in favor of more readily operationalizeable grammatical critera.

Another set of related problems arises when semantically *similar* processes have different syntactic realization. Consider, for example, the two clause types:

X *is like* Y
X *resembles* Y

Although a user might wish to place these similarly in the Upper Model, grammatically they are rather different. The former requires the grammatical features: {circumstantial-attribute, manner-participant, be-intensive}; the latter has features: {circumstantial-ascription, circumstantial-process}. In the present Nigel grammar, these are distinguished by the inquiry *circumstantial-ascription-q*, which would need to examine the Upper Model. Therefore, in order to obtain the differing syntactic structures a further distinction would need to be set up at the Upper Model level.

The realization class view therefore makes it difficult for users to formulate their input specification to the system unless they know precisely the form of linguistic expression that they require. Since the realizational link between Upper Model categories and Nigel has been made so tight for the very purposes of achieving readily describable criteria, it is sometimes (and increasingly once more users attempt more varied modelling) necessary to subordinate a concept in a counter-intuitive position simply in order for the language required to result. This certainly undermines the semantic integrity of the Upper Model as an interface ontology and moves the entire classification towards a less abstract level of information. It needs to be remembered, however, that only when the grammar is fixed, is a specific, determinate Upper Model required — furthermore, that Upper Model is even partially determined by the particular grammar that is specified.

Finally, one further problem with the interface ontology instantiated by the Penman Upper Model lies precisely in the simplicity of the relationship constructed betweeen domain model and Upper Model. We have seen that this is achieved by literally *classifying* (in the formal sense of adding into the subsumption lattice) domain model concepts in terms of the categories from the Upper Model. Following this operation, the Upper Model and the domain model form a single inheritance hierarchy and the domain concepts directly inherit the possibilities for surface realization defined for the Upper Model concepts. This operation is currently performed only once for each domain and, while simplifying input expressions, it means that the relationship between domain and Upper Model is not being handled particularly flexibly. In fact, once the classification is complete, the complete ontology can be interpreted as having collapsed into a mixed ontology of the type described for LILOG: both particular domain concepts and general linguistically motivated concepts occur in the same subsumption lattice. This treatment of the relationship between a potential conceptual ontology, containing detail knowledge of a domain, and the interface ontology, containing a semantic classification of possibilities for linguistic expression, needs to be made considerably more flexible to avoid the problems of mixed ontologies described both above and below.

3.3.2 Interface ontologies that are too abstract

Interface ontologies exhibiting the problem of being too abstract are more commonly found in small scale systems: the problem of not being able to specify the mapping down to grammar and lexis in a convenient and expandable form often prevents large-scale development from getting very far. Such projects (e.g., POLYGLOSS, ACORD and many others) begin by adopting classes of categories developed, for example, in analytical philosophy or natural language semantics — such as the event types of [Vendler, 1967], temporal categories such as those of [Moens and Steedman, 1988], the semantico-'conceptual' predicates proposed by [Jackendoff, 1983, Jackendoff, 1990], event structures of [Pustejovsky, 1988], and many others. As long as restricted grammatical possibilities are entertained, for example to enable research on particular focused areas of semantics-syntax, then such ontologies are adequate — even useful, since the focusing allows greater depth in the semantic account to be achieved. It should also be the case, however, that this work then feeds back into more general and broader ontology work, and this

happens much too rarely. It is also sometimes unclear what the relationship of these ontologies would be to a more abstract conceptual ontology — this may be expressed formally, for example, in terms of a model-semantic theory but the details are often left for future work.

3.3.3 Brief discussion

It is clear that ontologies of the interface type that are more closely bound to language are nevertheless most useful for NLP systems that want to deal with a wider variety of actual language phenomena. The increase in abstraction may not be so very great in comparison to a desired conceptual ontology, but it is nevertheless better than working with grammar and lexis directly. Such work is also much more likely to be stable in the face of changing theoretical positions and more justifiable with respect to actual linguistic data. It is, then, natural that one further type of NLP projects attacking the problem of large-scale ontology construction is that of 'dictionary'-oriented projects, such as EDR [Matsukawa and Yokota, 1991] and ACQUILEX [Calzolari, 1991]. The EDR project aims at producing a 'concept dictionary' containing 400,000 'word senses' for English and Japanese, and ACQUILEX is concerned with producing a re-usable 'lexical knowledge base' that classifies entries according to taxonomies of semantic categories and relations between those categories. Both projects have constructed sizeable semantic taxonomies relying strongly on differences in lexico-grammatical realization for the categories adopted. The taxonomy organization and categories found in ACQUILEX have similarities to the view of lexical semantics proposed by [Pustejovsky, 1991] where, again, *oppositions* in linguistic behavior are an essential motivating criterion. Another large project partly leading up to this work, and now related to the KBMT work mentioned above, was the MIT Lexicon Project where extensive classification of lexemes was undertaken on the basis of the differing grammatical patterns that the lexemes may enter into.

Although the construction of large knowledge bases at this level of abstraction is bound to offer a definite improvement in our ability to rely on linguistic motivations in future ontology design, their availability will not of itself bring about that design. It is still necessary to consider methodologies for using such information so that appropriate ontologies for general NLP use can be constructed. Therefore, in the next section I will relate the kinds of ontologies that we have seen in this section to compatible linguistic theories. Without a broader view of what is being done *linguistically* when categories for a particular kind of ontology are proposed, I believe it is unlikely that progress will be made. As long as the categories developed are sufficiently close to the surface details of language to remain objectively verifiable, i.e., remain in the realms of syntax and lexico-grammatically oriented interface ontologies, useful classifications can be constructed. For more abstract levels, however, the support of theory become crucial for defining methodologies, questions, and possible solutions.

4 Linguistic support (or otherwise) for the ontology types

In this section, I will follow the ordering of the discussion of ontology types of the previous section: i.e., first linguistic theories compatible with the design of mixed ontologies will be mentioned, followed by the kind of linguistic theory that is more supportive of distinct interface and conceptual ontologies. I will not raise the issue here of the relationship between 'conceptual ontologies' and possible linguistic theories, since one of the defining phrases that is often used about this level of abstraction is its very *extra-linguistic*ness. This does, however, depend on the view of the linguistic system that is adopted and I will mention something about this later. Finally in this section, I discuss some disadvantages of the former approaches when considered as a methodology for developing the kinds of resources necessary for NLP systems.

Before beginning the discussion, I should however briefly note the motivation for an exclusion of forms of semantics such as situation semantics, model-theoretic semantics of various kinds, etc. below. Such accounts are not immediately relevant to the discussion at hand precisely because they have not

been concerned with the construction of representations that are directly supportive of *ontologies*. That is, regardless of whether the formal account of semantics proposed in some particular framework contains sets of predicates that are of mixed ontological status, or are purely conceptual, or purely (linguistically) semantic, we find one crucial component of ontological engineering missing. Those categories are not typically built up into subsumption lattices of sorts sharing various general properties of use for further domain classification. It is clear that many of these theoretical approaches could easily move in this direction, and with the increased use of sorts in linguistic theory at all levels of description some first steps have been taken (e.g., [Sag and Pollard, 1991, p78], [Nerbonne, 1992]). However, as pointed out by [Onyshvekych and Nirenburg, 1991]:

> 'The crucial point is that in order to have an explanatory power, the atoms of [a] meaning representation language must be interpreted in terms of an independently motivated model of the world. Moreover, if any realistic experiments have to be performed with such an NLP system, this world model (sometimes called an *ontology*) must be actually built, not only defined algebraically.'

Therefore, until the problem of ontology construction on a realistic scale itself becomes an issue for an account, that account remains of less central concern for the current discussion.

4.1 Mixed ontologies and linguistic theory

The closest linguistic approaches to support mixed ontology design such as that found in LILOG are, perhaps surprisingly, those compatible with the work of [Jackendoff, 1983, Jackendoff, 1990]. Jackendoff adopts the position that the semantic level of representation with which he is concerned is also *conceptual*, i.e., common to modalities such as language and vision [Jackendoff, 1983]. As pointed out by [Herweg, 1991], approaches that directly link syntax with conceptual interpretation now occupy a rather standard position in mainstream linguistics and so there are many approaches that could be described. That of Jackendoff is probably one of the most developed and well known in this direction, although there are also similarities to be drawn with work in Cognitive Linguistics [Langacker, 1987, Talmy, 1987] and directions such as that of [Wierzbicka, 1988]. All of these approaches share an orientation to language as an instrument for revealing facets of conceptual organization. This is stated most clearly by Jackendoff in terms of what he terms the *Grammatical Constraint*:

> '... it would be perverse not to take as a working assumption that language is a relatively efficient and accurate encoding of the information it conveys. To give up this assumption is to refuse to look for systematicity in the relationship between syntax and semantics. A theory's deviations from efficient encoding must be rigorously justified, for what appears to be an irregular relationship between syntax and semantics may turn out merely to be a bad theory of one or the other.' [Jackendoff, 1983, p404]

Given his equation of semantic structure and conceptual structure, this becomes largely equivalent to statements such as the following describing the basic claim of of cognitive linguistics:

> "... across the spectrum of languages, the grammatical elements that are encountered, taken together, specify a crucial set of concepts. This set is highly restricted: only certain concepts appear in it, and not others... [This] set of grammatically specified notions collectively constitutes the fundamental conceptual structuring system of language. That is, this cross-linguistically select set of grammatically specified concepts provides the basic schematic framework for conceptual organization within the cognitive domain of language." [Talmy, 1987, p165/6]

This position also appears in the approach of Pustejovsky to the relation between lexemes and their interpretation in context; as he writes,

> 'The meaning of words should somehow reflect the deeper, conceptual structures in the system and the domain it operates in. This is tantamount to stating that the semantics of natural lan-

> guage should be the image of nonlinguistic conceptual principles (whatever their structure).' [Pustejovsky, 1991, p410]

These approaches are all described by the first architecture depicted in the diagram of Figure 1. Each suggests that there is a portion of the conceptual ontology that has a direct linguistic connection and that that portion should have just the same kind of organization as the rest of the conceptual ontology. A specification of the semantics of some expression is simultaneously a (possibly partial) specification of a conceptual specification. Again, this state of affairs receives a very explicit description from Jackendoff:

> 'This account of the syntax-semantics correspondence gives a principled account of the level of "argument structure" found in various versions of GB and LFG ... - a level of linguistic representation that lists the arguments of a verb, with or without their θ-roles. Such a list can now be simply constructed from the set of indices in the conceptual structure of the verb, and there is one index per syntactically expressed argument... In short, "argument structure" can be thought of as an abbreviation for the part of conceptual structure that is "visible" to the syntax.' [Jackendoff, 1983, p404/5]

By virtue of the Grammatical Constraint, therefore, Jackendoff adopts a very close binding of linguistic analysis and categories at his semantico-conceptual level of representation: available linguistic realizations and patternings lead directly to the positing of corresponding categories and relationships at the level of semantic/conceptual structure. In Jackendoff's case, the linguistic evidence admitted is organized in terms of $\overline{\text{X}}$-theory [Chomsky, 1980, Jackendoff, 1977] and so close correspondences appear between categories of this theory and categories of the semantic/conceptual structure. In particular, he states that:

1. "... every major phrasal constituent in the syntax of a sentence corresponds to a conceptual constituent that belongs to one of the major ontological categories."

2. "... the lexical head X of a major phrasal constituent corresponds to a function in conceptual structure — a chunk of the inner code with zero or more argument places that must be filled in order to form a complete conceptual constituent. The argument places are filled by the readings of the major phrasal constituents strictly subcategorized by X." [Jackendoff, 1983, p67]

Thus, he suggests the following approximation to conceptual structure for the sentence *The man put the book on the table* [Jackendoff, 1983, p68].

$$\left[\begin{array}{l}\text{EVENT}\\ \text{PUT}(\left[\begin{array}{l}\text{THING}\\ \text{THE MAN}\end{array}\right],\left[\begin{array}{l}\text{THING}\\ \text{THE BOOK}\end{array}\right]\\ ,\left[\begin{array}{l}\text{PLACE}\\ \text{ON}(\left[\begin{array}{l}\text{THING}\\ \text{THE TABLE}\end{array}\right])\end{array}\right])\end{array}\right]$$

This structure, if we ignore the textual information represented abbreviated here with THE, shows striking similarities with the input specification described earlier for Penman (cf. Figures 2 and 3). The structure may be glossed as stating that a predicate *put* of type *event* holds over three arguments: the first two are of type *thing*, the latter is an *on*-relation of type *place*. Each of the predicates are taken to be defined as semantico-conceptual categories motivated primarily by linguistic patterning. Further examples of the motivation of semantico-conceptual categories from linguistic evidence is the following list of example categories offered by Jackendoff:

Interrogative probe	*supports category:*
a. What did you buy?	[THING]
b. Where is my coat?	[PLACE]
c. Where did they go?	[DIRECTION]
d. What did you do?	[ACTION]
e. What happened next?	[EVENT]
f. How did you cook the eggs?	[MANNER]
g. How long was the fish?	[AMOUNT]

Subsequently, further categories of differentiations are made working from intuitions on the meanings of sentences and their constituents supported by example sentences. Moreoever, analogously to the perceived relationship between syntactic structures and rules for their well-formedness, Jackendoff takes the position that the inter-relationships between the semantic/conceptual categories will also be expressed in terms of well-formedness rules. An

example for the category [PATH] is as follows [Jackendoff, 1983, p166]:

$$[\text{PATH}] \rightarrow \left[Path \left\{ \begin{array}{l} \text{TO} \\ \text{FROM} \\ \text{TOWARD} \\ \text{AWAY-FROM} \\ \text{VIA} \end{array} \right\} \left(\left\{ \begin{array}{l} [_{Thing}\ \text{y}] \\ [_{Place}\ \text{y}] \end{array} \right\} \right) \right]$$

The combination of a number of rules such as these begins to define a hierarchy of interrelated categories analogous to the standard hierarchical organization that I have assumed appropriate for ontology construction.

A comparison of Jackendoff's semantico-conceptual categories with, for example, the superficially very different categories arising from cognitive linguistics is very illuminating concerning the role that motivations from language can play for ontology construction. The general methodology of proponents of cognitive linguistics is to examine 'grammatical' elements — however these come to be defined — in order to uncover the conceptual organization they presuppose. For example, Talmy offers the following break down of the *this/that* distinction in English.

> 'A closed-class element of this type specifies the location of an indicated object as being, in effect, on the speaker-side or the non-speaker-side of a conceptual partition drawn through space (or time or other qualitative dimension).' [Talmy, 1987, p168]

This is summarized as:

- a 'partition' that divides a space into 'regions'/'sides'
- the 'locatedness' (a particular relation) of a 'point' (or object idealizable as a point) 'within' a region
- (a side that is the) 'same as' or 'different from'
- a 'currently indicated' object and a 'currently communicating' entity.

By sampling across a wide range of languages the Cognitive Grammarian compiles a list of such distinctions and attempts to provide internal organization and structure rooted in a presumed linguistically relevant area of conceptual organization. The flavor of this organization can be seen in the following examples of proposed categories from Talmy.

Dimension "The category of 'dimension' has two principal member notions, 'space' and 'time'. The kind of entity that exists in space is — in respectively continuous or discrete form — 'matter' or 'objects'. The kind of entity existing in time is, correspondingly, 'action' or 'events'..." [Talmy, 1987, p174]. This is schematized as:

dimension	*continuous*	*discrete*
space :	matter	objects
time :	action	events

Plexity 'Plexity' is a generalization of notions such as singular and plural to cover actions also. For example:

		matter	*action*
a.	uniplex	*A* bird flew in.	He sighed *(once).*
b.	multiplex	Bird*s* flew in.	He *kept* sigh*ing.*

Boundedness 'Boundedness' is a generalization of notions such as mass and count with respect to nouns to include again actions in addition to objects. This Talmy relates to *imperfective* and *perfective* and similar terms in the treatment of verbs. Essentially, "[w]hen a quantity is specified as 'unbounded', it is conceived as continuing on indefinitely with no necessary characteristic of finiteness intrinsic to it. When a quantity is specified as 'bounded', it is conceived to be demarcated as an individual unit entity." ([Talmy, 1987, p178]). Similar, far more formal, expressions of this idea can now be found in a number of approaches (e.g. [Krifka, 1989]).

Dividedness "A quantity is 'discrete' (or 'particulate') if it is conceptualized as having breaks, or interruptions, through its composition. Otherwise, the quantity is conceptualized as 'continuous'." [Talmy, 1987, p180]

These categories hold of a given 'quantity' simultaneously and so classify that quantity along the dimensions described. Moreover, different linguistic consequences are intended to follow from each distinction. Although there are many interesting distinctions suggested which could help enrich proposed ontologies along a number of dimensions, the lack of an accepted, detailed grammatical framework nevertheless limits the generalizations that can be found. Langacker claims that:

> "...basic grammatical categories such as **noun**, **verb**, **adjective**, and **adverb** are semantically definable. The

> entities referred to as nouns, verbs, etc. are symbolic units, each with a semantic and a phonological pole, but it is the former that determines the categorization. All members of a given class share fundamental semantic properties, and their semantic poles thus instantiate a single abstract schema subject to reasonably explicit characterization. A noun, for example, is a symbolic structure whose semantic pole instantiates the schema [THING]... In a similar fashion, a verb is said to designate a **process**, whereas adjectives and adverbs designate different kinds of **atemporal relations**." [Langacker, 1987, p189]

Although with the proposed conceptual categories restricted in this way to follow from grammatical categories that are so directly 'observable', i.e., often inflectional and word-based such as singular and plural, mass and count, nouns and verbs, etc., one would not expect a particularly rich ontology, in fact, a large number of finely differentiated categories are set up — primarily on the basis of contrastive examples that do *not* rely on detailed syntactic analysis. This shows conclusively the value of examining a very wide range of natural occuring examples, in contrast to the oft criticised (e.g., [Rohrer, 1986]), but nevertheless still prominent, tendency in mainstream linguistics to study constructed examples in areas that illuminate the currently fashionable linguistic phenomena. Nevertheless, the lack of a formally specifiable mapping between the categories proposed and linguistic realization renders the consequences of establishing any particular set of categories almost impossible to investigate and this is certainly less of a problem in a contrasting account such as that of Jackendoff where the relation to a detailed account of grammar and lexis is always clear. The value noted above of being able to *test out* and justify proposed categories for ontologies formally applies here strongly. Jackendoff is able, therefore, even on the basis of rather limited linguistic breadth of motivation, to suggest a more detailed set of categories and interrelationships. The semantico-conceptual representations are substantially more abstract than syntactic classes (as evidenced by the generalizations that they permit to be drawn) but are nevertheless tied reasonably precisely with possibilities for linguistic expression. An ideal situation would therefore be to have a very broad, detailed and formally specified grammar, capable of describing very fine-grained grammatical and lexical differences.

Even despite the lack of formally specified mappings to linguistic form within cognitive linguistics, there has still been at least one significant application of its proposed concepts in a computationally context. This is in their use to provide a system of *semantic features* for stating meanings to be preserved across languages in machine translation [Zelinsky-Wibbelt, 1987, Zelinsky-Wibbelt, 1988]. Although the work suffers from the lack of explicit definition that the conceptual categories have so far received — making it difficult for coders using the semantic features reliably to classify the meanings that are involved — this situation may be improved significantly by some current work in progress[18] which is intended to improve the necessary connection between the semantic categories and their linguistic realization. The situation applying Jackendoff's categories in a computational context has, as would be expected, been more straightforward. A number of proposals have been made for such an application, and some have been implemented. For example, [Meteer, 1988] comments on the possible organization of abstract linguistic terms at the text message level for the sentence generator Mumble-86 that a system such as Jackendoff's could provide and we have already seen that both [Dorr, 1987, Dorr, 1990]'s work on the UNITRAN translation system and approaches within KBMT [Nirenburg and Levin, 1991] have implemented aspects of the semantico-conceptual structure.[19]

[18] For example by Cornelia Zelinsky-Wibbelt and Wiebke Ramm of IAI/Eurotra-D on syntactic tests that coders could apply to resolve difficult cases.

[19] Further analogous areas of research which often fall somewhere between the explicit grammatical foundation attempted by Jackendoff and the, until now, more impressionistic linguistic motivations of Langacker and others, include the large body of work on the 'conceptualization' and linguistic expression of spatial-temporal information, e.g.: [Herskovits, 1986, Bierwisch and Lang, 1989] and many others.

4.2 Interface ontologies and linguistic theory

In contrast to the accounts of the previous section, the separation of information found in the interface ontology and a more abstract conceptual ontology is consonant with theoretical positions that assume a higher degree of *stratification* in the linguistic system. The mixed ontology view goes well with a standard syntax-semantics-pragmatics distinction where 'semantics' includes the conceptual representation and 'pragmatics' provides procedures that operate over the semantico-conceptual representation to produce active interpretations in context. In this sense, pragmatics is not a further stratum in a linguistic system and has a distinct theoretical status to that of syntax or semantics. In contrast to this, the interface ontology architecture suggests at least a three-way stratification between lexico-grammatical information, semantic information, and a contextualizing level of 'conceptual' information. Each of these strata appears to have rather similar formal properties: most of the information of each, for example, would appear to be representable as a subsumption lattice defined over sorts, possibly augmented with structured inheritance.[20]

I have already mentioned one view of the linguistic system that seems compatible with this stratification: the approach to semantics and context proposed by [Bierwisch, 1982, Bierwisch and Lang, 1989] that acted as one influence for the LILOG design — even though the final specification of the ontology within LILOG does not seem to have remained in the spirit of this theory. Within the linguistic model of Bierwisch, conceptual representations are maintained strictly separate from semantic representations, and semantic expressions are used to constrain construction of conceptual expressions during interpretation. Thus, 'words' (actually lexicogrammatical patterns) are related to semantic forms which determine functions from contexts to conceptual structures. The distinction between the two levels in this kind of **two-level semantics** is nicely summarized by Michael Herweg as follows:

> 'Semantic representations are structured configurations of semantic units which, on the one hand, are determined by the grammatical system of the language in question and, on the other hand, are grounded in — or motivated by — the conceptual system. ... Conceptual representations are structured configurations of conceptual units, which are mental representations of certain aspects of the external world.' [Herweg, 1991, p152/3]

The two classes of categories — the semantic and the conceptual — thus have very different theoretical statuses and allow very different kinds of motivations. This is therefore precisely the kind of structuring of the linguistic system that one requires to support the use of interface and conceptual type ontologies.

The most successful of the interface ontologies described in Section 3.3, the Penman Upper Model, clearly has a natural relation to the stratification found in this kind of 'two-level semantics'. For example, the sorts of the Upper Model are determined by the grammatical system (concretely, the Nigel grammar component of the Penman system) as is required. Although there is also a relationship to be drawn with accounts that are explicitly seeking *semantic* organizations closely linked to language regardless of these organizations' further embedding at higher levels of abstraction,[21], the relation between a proper view of the Upper Model and two-level semantics becomes even closer when we examine instead of the Upper Model, rather the *theoretical position* of which the Upper Model is only a very partial instantiation: i.e., that of systemic-functional theory [Halliday, 1961, Halliday, 1978,
Matthiessen and Halliday, forthcoming].

Systemic-functional theory is a highly stratified general linguistic theory with respect to which the Penman text generation and its descendents have been, and continue to be, developed. In some ways perhaps analogously to the situation in LILOG, many aspects of the current implementation of the Penman system are not accurate instantiations of that theory. Of particular importance here is the very instantiation of the concept of linguistic strata

[20] Current work in information-based syntax makes this point for syntax [Pollard and Sag, 1987].

[21] Or, alternatively, by seeking an embedding in an account such as model-theoretic semantics to 'bottom out' in a formally specifiable way.

— since it is precisely this construct which is necessary for motivating the kind of multi-levelled representation that we find in interface ontologies and their contextualizing conceptual ontologies.

The notion of stratification in systemic-functional theory is depicted diagrammatically in Figure 5. The linguistic system is broken down here into three strata: lexicogrammar, semantics, and context. Between each stratum the same relationship — that of *realization* — holds. Systemic-functional theory is essentially a *functional* theory, i.e., one that is concerned crucially with the functions that language fulfills in particular contexts, and this informs the understanding of the realization relationship between strata as follows. Each higher-level (i.e., more abstract) stratum is seen as providing the *functional motivation* for the next lower-level stratum; and each lower-level stratum is seen as providing a resource that *generalizes* across the possibilities of the next-higher stratum [Halliday, 1978]. This gives us a more detailed view on how strata in the linguistic system interact than that usually found in stratified accounts. Additionally, each higher-level stratum is seen as *contextualizing* the levels beneath.

The organization of the Penman-style architecture version of systemic theory instantiates the stratification as follows. Nearest the surface there are *realization statements* of syntagmatic organization, or syntactic form. These statements are classified in terms of their potential for expressing communicative functions that are realized grammatically, such as asserting/questioning/ordering, active/passive, etc.: this denotes paradigmatic organization and is represented in terms of a *grammatical system network*. This organization captures the possible alternatives that are available given any choices that have already been made; i.e., a collection of 'paradigms' of the form 'a linguistic unit of type A is either an A of functional subtype X, or an A of functional subtype Y, ..., or an A of functional subtype Z' are given. At each level these subtypes are disjoint and serve to successively classify linguistic units along ever more finely discriminated dimensions. This formulation of classifications in terms of increasingly fine discrimination is in systemic-functional linguistics termed the principle of *delicacy*. The grammatical communicative functions are then in turn *motivated* by semantic distinctions that classify semantic circumstances according to the grammatical features which are appropriate to express those situations: this classification is the combined responsibility of *choosers* and *inquiries* [Mann, 1983]. Finally, the possibilities for classification that the inquiries have are defined in terms of the abstract ontology of the Upper Model. In relation to Figure 5, then, the Penman-style architecture represents a computational instantiation only for the *lower* two strata and the relationship between them.

While at a rather general level very similar to the breakdown proposed by Bierwisch, the systemic-functional account also goes into more detail about the internal organization of each stratum. It is this feature which is largely responsible both for the more abstract status that has been achieved for the sorts of the Upper Model and for the early adoption of the principle of motivating sources on the basis of the grammar. Not only is all grammatical variation captured by abstract choices between minimal grammatical alternatives, but also all such abstract choices must have explicit motivations, or semantic conditions, defined. Only then is the grammar fully defined as a resource for grammatical expression: we have to know what each grammatical possibility is an expression of. This has naturally given rise to the notion of *covering the grammar* in terms of a set of motivations for each choice that the grammar offers. This is depicted graphically in Figure 6. The categories necessary for this motivational covering are then organized into sorts in a subsumption lattice — thus defining the Upper Model.

It is worth noting that this provides a very strong methodology for interface ontology construction. Until a grammar alternation is explicitly connected into a motivational relationship, the alternation is considered to be only formally (in the sense of *linguistic* form) defined. The grammar in fact acts as a (highly structured) list of phenomena that require semantic motivation. In addition, the *functional* organization of the grammar itself goes a long way towards providing a useful pre-classification of syntactic phenomena so as to be amenable to systematic semantic interpretation. The extra boost in abstraction

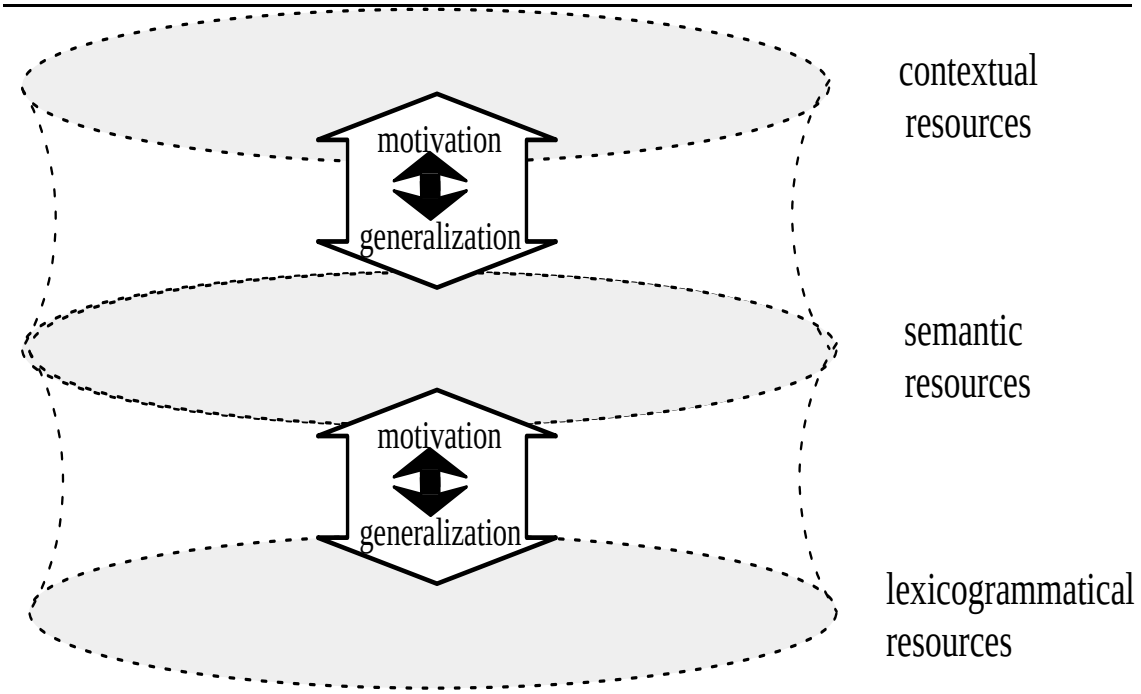

Figure 5: Stratification with systemic-functional linguistics

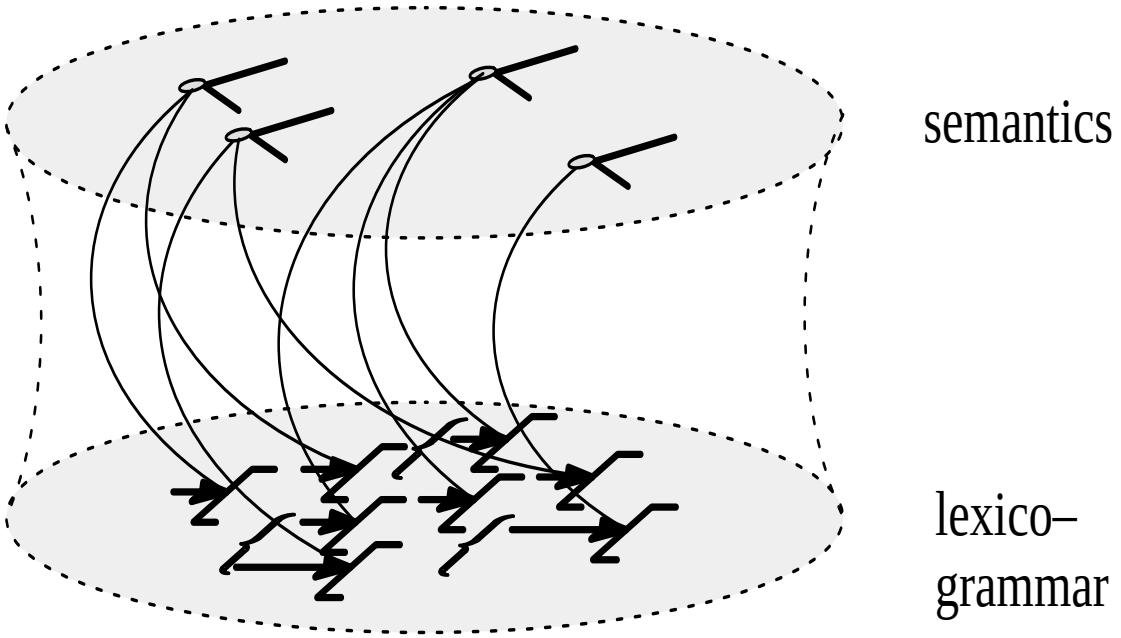

Figure 6: Covering the grammar semantically

that the grammar offers is responsible for the increased level of abstraction that the Upper Model has already achieved.

I have noted, however, that the Upper Model does not instantiate the full organization required by the theory. Some of the consequences of this have already been mentioned. For example, the *upwards* relations of the Upper Model to context has not been modelled within Penman in terms of a 'realization' relationship: domain models are directly subordinated to the Upper Model hierarchy. I will return to this and some other problems below. Importantly, since the full input of the theory has not yet been taken into account, we have available a number of possible directions for development that may provide a far more sophisticated implementation of both interface and conceptual ontologies.

4.3 A comparative evaluation

In this section, I will apply the possible linguistic theoretical underpinnings that could be provided for the different ontology types in order to consider those ontology types more critically. I will suggest here that there are clear reasons for dispreferring accounts that adopt a mixed ontology approach. Subsequently, in the next section, I will discuss possible future developments for interface and conceptual ontologies drawing further on the connections to theory established.

We have seen that the type of account based on a style of argumentation such as that of Jackendoff manages to gain abstractness while still maintaining contact with details of linguistic realization. I have noted that the increase in abstraction is a generally necessary property for improving the functionality of NLP systems. One of the principle differences between such an account and the linguistic theories supportive of interface ontologies was in the degree and explicitness of stratification. One can ask the question, therefore, is there any evidence for the more stratified view of the linguistic system? If it proves necessary, or beneficial, to differentiate between information that is particularly linguistic and the kind of information sought in accounts of real-world, commonsense knowledge, then a mixed ontology will not be sensitive to this. A very important issue to address is, therefore, whether the selection between a mixed ontology and a more differientiated set of interrelated ontologies is one which is still open to debate, or arbitrary — or are there grounds for deciding for one architecture over the other.

4.3.1 Populating a mixed ontology

If we assume that we have an account such as that proposed by Jackendoff, possibly augmented by a range of concepts from cognitive linguistics with a more formally expressed relationship to the lexicogrammar, it is still the case that there the resulting ontology is not yet very large. The number of general sorts that occur in, for example, [Jackendoff, 1990] (i.e., not the conceptual equivalents of lexical items, which seem to be introduced freely), is less than 40: these include predicates such as EVENT, STATE, BE, ORIENT, PATH, GO, WITH, FROM, TO, TOWARDS, INCHoative, REACT, AFFect, etc. Most conceptual items are decomposed into these 'primitives'. I will not discuss whether or not these items are good candidates psychologically for conceptual primitives, but relying on this small set of categories is unlikely to capture many generalizations of linguistic expression when a broad lexicogrammar is considered. If we include experience such as that obtained within the development of the LILOG ontology or the Upper Model, many intermediate sorts will express useful generalizations over distinct linguistic patterns. Relying on a smaller than necessary set of categories either misses generalizations or places more work on the mapping with lexicogrammatical form. The methodological question arises of how the sort hierarchy is to be extended beyond the very general categories that most attempts at ontology construction assume as basic on intuitive grounds.

The primary source of evidence for extension is the classification of lexicogrammatical patterns. This posits semantic features that co-occur with particular classes of Lexical Conceptual Structures. But these classes are constructed on the purely *syntactic* linguistic behaviour of the investigated lexemes. This, while being the best methodology available and one I have defended throughout this paper, cannot itself be expected to give rise to *conceptual* classes. Only the assumption that such semantic patterns are simultane-

ously conceptual makes this plausible: there is no obvious connection to be drawn between aspects of domain and commonsense knowledge and lexically derived categories. The latter are often subject to criticism for being too shallow even for an interface ontology: they must appear very unlikely candidates for a conceptual ontology. As we have found with the problems with the Upper Model (Section 3.3.1), there is no guarantee that particular domain-motivated categories will choose lexically-motivated categories that belong to a consistent more general ontological type. More often items belonging to very different lexical classes are treated as semantically equivalent for speakers' expressive purposes. Representing this in a single ontology then requires that concepts may be consistently classified along the two dimensions simultaneously: which complicates the formal properties of the resulting ontology considerably since exactly what may be inherited where becomes unclear.

This is shown concretely in the linguistically motivated evaluation that [Lang, 1991] undertakes for LILOG. There he examines the sorts proposed for the ontology according to the kinds of motivations accepted for their inclusion. He finds the following differentially motivated sorts all combined in the single subsumption lattice:

- 'Conceptually based sorts' which are included on extra-linguistic (conceptual) grounds.
- 'Text base specific sorts' which are concepts corresponding to special vocabulary items required by the particular domain and text with which LILOG as a project was concerned.
- 'Sorts projected from the grammar' which are notions found in the grammar, such as *preposition*, transferred to the ontology.
- 'Sorts of mixed origin' which are concepts where both extra-linguistic and linguistic criteria are involved.

This mixing of motivations organizes itself loosely according to the vertical and horizontal dimensions in the hierarchy. Thus,

> 'The vertical structure of the sort hierarchy, which is based on the subsumption relation, draws mainly on the availability of corresponding linguistic labels categorized as nouns ... or as verbs ... However, the horizontal dimension of the sort hierarchy, that is the selection of subsorts to be assigned to a common supersort, is mainly determined by features that emerge from our extra-linguistic conceptual knowledge of objects and spatio-temporally specifiable events or situations ...' [Lang, 1991, p466/7]

Lang shows the following problems for the resulting organization in a single subsumption lattice that this inconsistency, or variety, of motivations for concepts in the hierarchy creates. First, since extra-linguistic or conceptual criteria are less than well understood, there is a degree of arbitrariness in the categorizations that appear. Second, it is never clear from the concepts that are found in the hierarchy alone whether they are to be expected to have a corresponding linguistic effect or not. Third, the co-existence of distinct kinds of concepts means that the precise meaning of 'subsumption' with respect to particular cases is underspecified — different kinds of concepts have different relations between their 'wholes' and their 'parts' and until this is clarified it is unclear what kind of subsumption actually holds. These differences entail different formal properties so that different objects can call for different inheritance properties. Thus, for example, a supposedly general 'part-whole' relation is intended sometimes as 'is a component of', sometimes as 'is spatially included in', somtimes as 'pertains to', sometimes as 'inalienable possession', sometimes as 'alienable possession', etc. This range of possibilities makes the inferences that in fact follow from any statement in the ontology far more difficult to foresee and substantially complicates in any case any axioms for inference that are designed.

This can also be seen concretely in many versions of semantics where a mixed ontology is relied upon — in order to handle the very flexibility of the relationship between the concepts that are to function for the linguistic expression, and those which are not, complex and often unconstrained mechanisms are introduced: the

'projective inheritance' of [Pustejovsky, 1991] and many instances of 'type coercion' as used by, e.g., [Sag and Pollard, 1991, Pustejovsky, 1991] are probably prime examples of this, but there are many others. A mixed ontology is, therefore, a very weakly constraining theoretical construct, which does not provide optimal assistance either for theory construction or for system construction.

4.3.2 Stratification

The just mentioned flexibility of relationship between conceptual categories and the categories that are determinative of their linguistic realization is a very typical property of a relationship between linguistic strata. It is this very flexibility, in fact, which provides the primary linguistic evidence for stratification. As an example of this, consider the following issue of ontological design.

Regardless of whether a mixed ontology is adopted or not, some portion of some ontology is assumed which offers an expression of the chunking that language expects and demands of knowledge if it going to be expressible through the grammatical and lexical resources of the linguistic system. One question that can be asked, therefore, is: is the information in a conceptual ontology that will support this chunking *already organized in this way or not?* If it is then it will be straightforward to construct a mechanism such as that suggested by Jackendoff above, whereby one simply 'takes a view' on some conceptual structure and already has a specification of the semantic predicate-argument structure which can in turn control the grammar and lexis to produce appropriate results. If not, however, then some reorganization of the structure will be necessary. In all examples that are presented of alleged conceptual structures that are already appropriate for direct lexicogrammatical realization, e.g., by viewing as predicate-argument structure, we can make the following observations:

First, the lexical items and class of grammatical patterns appropriate is already so highly constrained as to follow directly from the expression; for example,

[State ORIENT ([Thing WEATHERVANE], [Path NORTH])]

as one reading for the sentence: *The weathervane pointed north* [Jackendoff, 1990, p74]. Certain variability in lexicogrammatical expression will be produced by the mapping rules of syntax formation, but other decisions, including: the choice of word for the concept *weathervane* given that the hearer might not know what a weathervane is, or that the sentence may be uttered among world-experts on the subject of weathervanes who would normally select a far more restrictive description, etc. have already been built into the description. Widely differing selections of possible expression according to text type, register, formality, situation, time availability (cf. [Hovy, 1988, Bateman and Paris, 1]) are excluded.

Second, the *granularity* of the corresponding language has also been built into the description. For example, we know that a sentence is going to be produced (or if the linking rules are good enough: a sentence or a nominalization) rather than a short discussion of the wind's effect on an object whose position of equilibrium under the pressure of the wind serves as an indication of the wind's direction. A nice example[22] of maximal flexibility here might be the difference, for example, in the language produced in response to the conceptual real-world category *beer* for the purposes of a dictionary entry, e.g.,

> '**Beer** is a bitter alcoholic drink made from grain. There are a lot of different kinds of beer.' [Collins COBUILD English Language Dictionary, 1987]

and that produced for the purpose for an entry in an industrial chemical encyclopedia, which goes on for 40 pages.

The response to both of these problems within the semantico-conceptual approach is straightforward: the differences are expressed beforehand in the semantico-conceptual organization and are produced by conceptual processes for information organization and management. But this misses the generalization that *regardless* of the information to be expressed that same *linguistic* granularity is imposed: there will be a set of descriptions of some predicate with an argument structure, including specifications of participants and circumstances. The two sentences concerning beer in the dictionary and the hundreds in the

[22] Due to Karin Haenelt.

encyclopedia all exhibit the same kind of organization. Knowledge is variable scale, but language is predominantly fixed-grain,[23] as defined by the grammar. This means that for all the knowledge available in the semantico-conceptual ontology, there need to be construction mechanisms available which convert some selected fragment of the information, of any scale, and produce an appropriate sized chunk of semantico-conceptual structure for motivating a sentence.

With unconstrained inferencing across the knowledge base this may be achievable by inheriting constraints back from the grammar and checking the equivalence of constructed semantico-conceptual structures with the originally selected fragment. But, crucially, for all such selected fragments, the same class of 'semantico-conceptual' paraphrases will be potentially available: i.e., those licensed by their grammatical expressability. Furthermore, also regardless of the originally selected semantico-conceptual fragment, the lexico-grammatically licensed set of 'semantico-conceptual' specifications govern specifiable sets of inferences that operate *only on such* specifications: for example the inferences that determine the textual variations that are appropriate when realizing the specification lexicogrammatically [Bateman and Matthiessen, to appear], that certain abstract semantic classifications apply for which there is no conceptual evidence [Schriefers, 1990], and others. Thus, not representing this distinguished set separately fails to capture a significant generalization about the organization of the linguistic system as a whole.[24]

Finally, it is worth emphasizing that this flexibility between strata is typical and not unique to the relation between semantics and conceptual levels of representation. The relationship between, for example, the Upper Model and the lexicogrammar already exhibits much of the same kind of flexibility. For example, the expressive resources of the grammar of nominal groups is not restricted to the single grain-size of sorts that are subtypes of an Upper Model sort *object*. It is equally possible to realize Upper Model classified *events* as nominal groups or configurations of events as single clauses if the textual conditions are appropriate. [Bateman and Paris, 1] present other examples of this theoretical flexibility for other categories in the grammar. It is not at all surprising given the theoretical similarity, therefore, to find exactly this kind of flexibility again between the sort lattice of the semantic ontology and that of the conceptual ontology.

This discussion of stratification is summarized in Figure 7. Here we see three strata and the repeated variability in expression that any selected *semantic* specification has. Crucially, the common, reoccuring coding possibilities that are available for all elements from the conceptual stratum are not repeated at that level, but are factored into a single statement at the level of the semantic interface with a mapping from sorts at the conceptual stratum to sorts at the interface stratum. Not representing this generalization both guarantees a complication of the theory and makes a usable NLP system based on the theory unlikely. Again, the power of the theory to bring methodological and contentful constraints to bear on system design is compromised.

5 Some Principles and Methods; and some former puzzles resolved

The discussion up to this point has attacked mixed ontologies on the basis that they are internally inconsistent, and has criticised the non-statificational linguistic accounts underlying such mixed ontologies on the basis that

[23] Apart from the resources for combining clauses, nominal groups, etc. into 'complexes', which are not relevant to the current argument.

[24] It is also engenders dubious NLP system design; factoring out the commonalities in a separate stratum is analogous to the following application of object-oriented programming:

> "In an object-oriented application ... the system uses predefined mappings from objects to the routines that know how to process those objects (or can choose among different routines depending on the context). The efficiency of using predefined mappings for known types comes in drastically reducing or entirely eliminating search; the onus is put on the developer to define the decisions available to a type at each level, rather than presenting all options at all times and letting a search procedure find the best one." [Meteer, 1989, p6]

This is also one property of using an interface ontology such as Meteer's or the Upper Model.

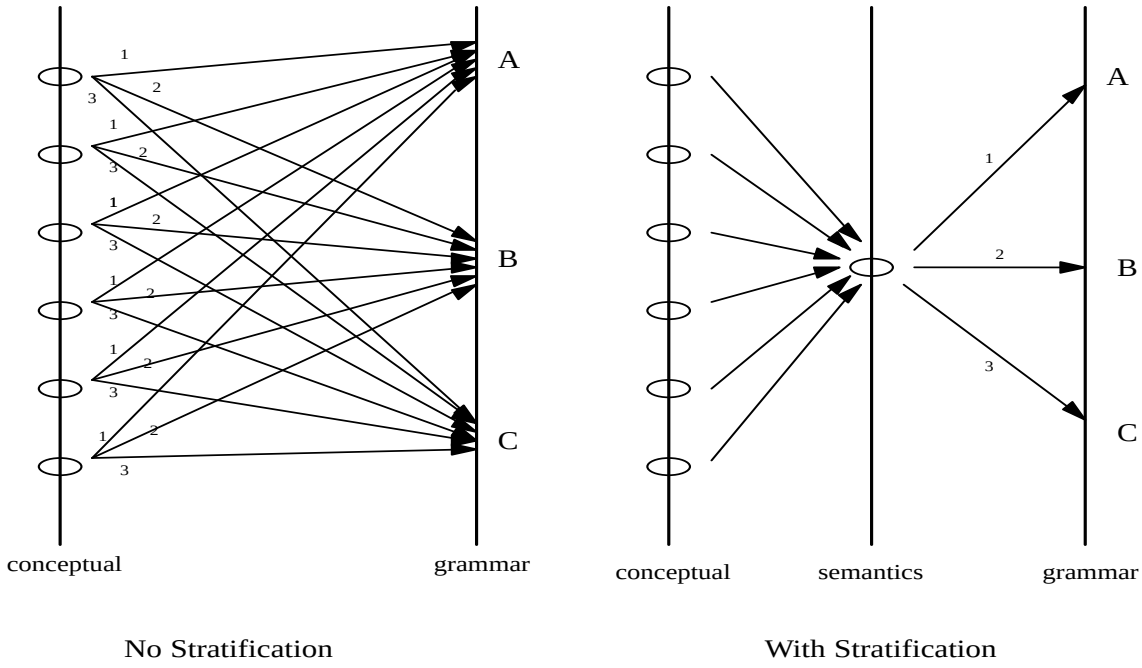

Figure 7: Capturing generalizations via stratification

they fail to capture theoretically important and practically useful generalizations. Both weaknesses have one consequence in common: they provide a seriously reduced set of constraints for ontology design and construction. Since one purpose of this paper is to suggest principles and methodologies for ontology construction, mixed ontologies and understratified linguistic accounts are clearly to be avoided. The kinds of ontologies most appropriate for NLP systems and for which linguistic support needs to be sought can now be restricted to the following two types.

- Type O_i: an abstract semantic organization underlying our use of grammar and lexis that is motivated on essentially *linguistic* grounds and that acts as a complex interface between lexicogrammatical resources and higher-level strata in the linguistic system — the categories of this interface should be maximally general, i.e., apply across distinct real-world situations, but specific enough to maximally constrain possible lexicogrammatical expression.
- Type O_c: an abstract organization of real-world knowledge (commonsense or otherwise) that relates downwards to the interface to lexicogrammar.

With these restrictions in place, I will now go further and suggest some particular guidelines for ontology construction. In order to do this, however, it is also necessary to make some further commitments to the kinds of information that will be made available at particular levels in the linguistic system. The reason for this is that the linguistic, and particularly lexico-grammatical, constructs are essential for guiding ontology design. This follows the increasingly wide range of linguistic theories that are returning to the position that the relation between grammar and semantics is not arbitrary; we saw a selection of these in Section 4 above, e.g., [Langacker, 1987, Talmy, 1987, Wierzbicka, 1988, Jackendoff, 1983, Halliday, 1978]. If we accept this, then it is also to be accepted that the selection of *particular accounts of lexicogrammar* has consequences for the subsequent ontology design. Since such consequences cannot be avoided, it makes sense to make selection decisions in ways which will maximally help in the task of ontology construction overall. I will distinguish between

decisions in the following two areas:

- type of grammar
- contents of grammar

and will make some firm suggestions for the former and discuss the consequences of differences that arise in the latter.

First, we can note that the most successful interface ontology developed so far is probably that of the Upper Model. The Upper Model has achieved both a detailed account and a generally applicable account. We can ask, then, what is it about its underlying theoretical organization that let this occur? Second, we can further note that although the Upper Model is the most detailed instantiation of an ontology of type O_i that has been developed, is nevertheless not a full instantiation of the theory on which it is based. It is therefore worthwhile considering briefly what additional constraints the theory could bring to bear if it were to be more fully implemented.

The kind of grammar on which the abstractions proposed in the Penman Upper Model is easily classified. It is a *paradigmatic-functional* grammar exhibiting the standard Hallidayan *metafunctional* diversification [Halliday, 1985, Matthiessen, 1990, Matthiessen and Bateman, 1991].
This means that it is organized, firstly around *choice* — the paradigms of grammatical constructions that stand in functional opposition — and second around a factorization of that choice according to its semantic motivation: is the choice to do with the propositional content of the linguistic entity to be classified (*ideational*), is the choice to do with the textual placing of the linguistic entity to be classified (*textual*), or is the choice to be classified as to do with the interpersonal relationship between speaker and hearer or with the attitude of the speaker towards the information expressed by the linguistic entity (*interpersonal*). The motivations for the choices provides hypotheses concerning the sorts necessary for controlling those choices. The Upper Model has been derived by considering motivations for those choices *exclusively assigned to the ideational metafunction*: there is no mixing of categories across metafunctional domains.[25]

This builds into the design of an ontology motivated in such a way as the Upper Model the following features.

First, we require an ontology that is significantly more *abstract* than syntactic realization classes. I have already suggested how this has been achieved with the Upper Model. The grammar, being organized in terms of a functional classification of possible constraints on constituency structure, is *already* more abstract than constituency structure *per se*. Further classification across the paradigms uncovered is then automatically more abstract and achieves a generalization across particular lexico-grammatical contexts that supports a greater flexibility of expression of input expressions. The strict relationship to the grammatical stratum also makes sure that the kinds of mixed sorts that [Lang, 1991] finds and criticises in the LILOG hierarchy *cannot* occur: either an (interface, i.e., semantic) ontological category has a specified consequence for lexico-grammatical expression or it is not accepted.

Second, given the stratification suggested by the theory the Upper Model is automatically only the 'next level up' in the linguistic system: it is an ontology strongly connected to grammar below. It does not, by itself, provide the necessary organization of higher level ontologies. Thus, in short, we see that an organization closely reminiscent of a two-level semantics is automatically achieved, and that *both* levels require ontologies.

Third, we have seen that it is a design goal that an ontology be as general as possible — that it helps with classification across domains, tasks, and applications, but also be substantial enough to provide a rich scaffolding for domain description. This raises the question: How can we guarantee that a proposed ontology is as general as we require? We can now see that ontologies such as the Upper Model, which are based on motivations from grammar, are *guaranteed* to have the domain-independence required

[25] Although it is perfectly possible to imagine applying the same 'grammar-as-filter' methodology on underlying motivational ontologies as carried out for the ideational metafunction — cf. [Bateman, 1991] for examples of this applied to the textual metafunction and [Matthiessen and Bateman, 1991] for general discussion — the resulting organizations of information have very different properties.

of them. Since ontological categories are motivated by the grammatical distinctions (and not by more arbitrary lexical collections found in a given domain), those categories are forced to be *at least* as general as the grammatical categories. It is, therefore, very unlikely that [Klose and von Luck, 1991, p462]'s claim that the LILOG ontology has a 'more domain-independent status' than the early version of the Upper Model described in [Mann *et al.*, 1985] would apply to current versions of the Upper Model.

But we can go further and move beyond the kind of generalizeability that refers simply to domain-independence — which is generalizeability 'upwards' in the linguistic system, and beyond generalizeablity across the lexicogrammar — which is generalizeability 'downwards' in the linguistic system. When we also consider the *metafunctional* organization of the linguistic system posited by systemic-functional theory, then we can see that generalizations both across 'text instances' and across 'speech functions' are also guaranteed — i.e., generalizations 'horizontally' across the same stratum of the linguistic system. This is depicted graphically in Figure 8. These constraints rule out certain other potential sorts from the ontology, e.g., sorts concerned with the particular appearance of an entity at a given position in a text or with the speaker's attitude towards an event. Certain of the sorts found in [Meteer, 1989]'s interface ontology are good examples of the former kind. Having such sorts requires reclassification of domain information whenever a domain object is used in a text, since the textual statuses of domain objects changes over the development of a text — i.e., from new to given, from theme to rheme, etc. This change of course needs to be represented: the point is that representing such information in the interface ontology again *mixes* very different kinds of information — although this time on a 'horizontal' dimension across the linguistic system rather than a 'vertical' one.

The kind of grammar that we employed as the initial motivation for guiding the development of the Upper Model has, therefore, gone a long way towards ensuring that the properties desired of ontologies obtain. But an area of flexibility in the description then arises from the *depth* of grammatical description, i.e., the *contents*, rather than the type. Particularly within the systemic-functional approach, lexical descriptions are seen as more specific versions of grammatical descriptions — there is no difference in *kind*. Thus, if we push lexicogrammatical description further in the direction of lexis, we automatically push further the depth of motivating semantic ontology constructs that are needed. This bifurcation in potential description needs more theoretical work before we can make any firm statements about whether it is more helpful to pursue one at the expense of the other, or whether they should be pursued in parallel as has been the case with the more general area of the grammar.

We can now also consider some possible improvements and explanations for some awkward phenomena/intuitions that have previously hindered ontological engineering. For example, if there is a stratification of the kind argued for, why is it that suggestions for conceptual structure that have been put forward in a number of approaches appear also to be candidates for representation as sorts in the interface ontology? — When the categories of the Upper Model, for example, are examined, many similar classes to the proposed 'conceptual' ontology work are to be found.

To give a concrete example of this, [Lang, 1991, p474], after careful discussion concerning the problems of a mixed ontology, defines some basic assumptions concerning the structure of the conceptual ontology drawn from earlier work, including [Bierwisch and Lang, 1989]. With respect to these assumptions, he outlines the following set of conceptual **domains** which are to form basic subsorts of the conceptual ontology:

D_1: objects; D_2: substances; D_3: locations;
D_4: time intervals; D_5: events; D_6: attitudes

We can also note here similarities with some of the classes above from [Jackendoff, 1983, Jackendoff, 1990, Langacker, 1987, Talmy, 1987], etc. But these are also sorts already found, for example, in the Penman Upper Model, where they have been entered purely on the grounds that they are necessary to directly constrain possible grammatical realizations. Is it the case that the claim we saw above by [Gust, 1991, p133] that: 'there are continuous variations between semantic forms and conceptual structures' is, after all, true? Can we introduce strict stratification and still

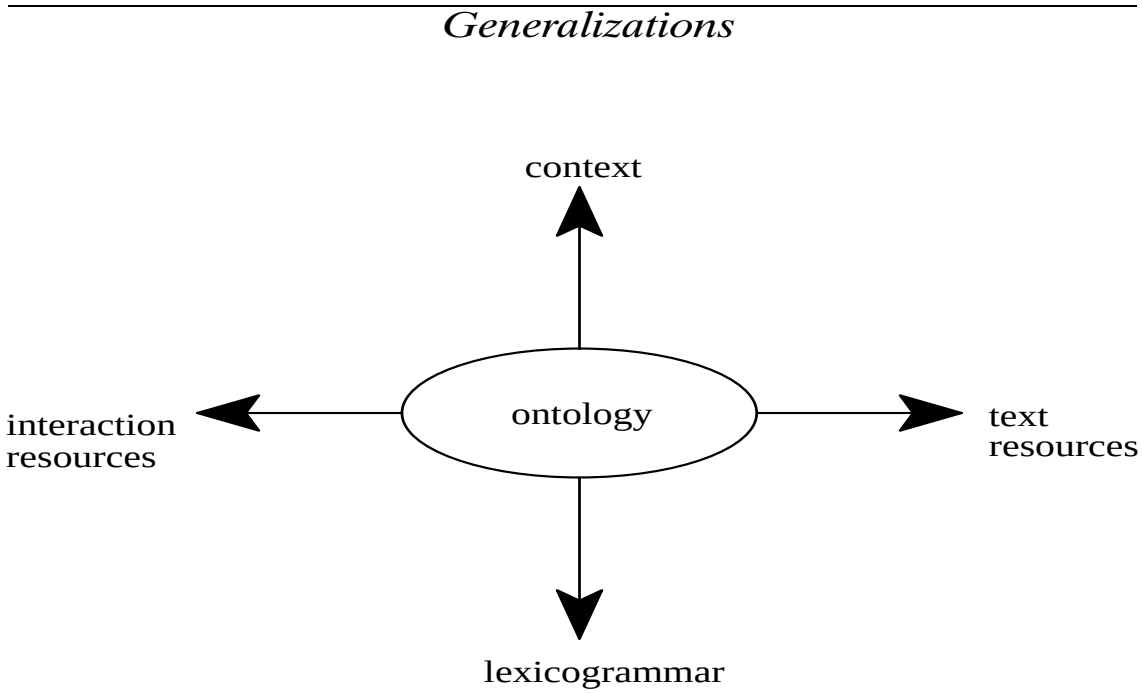

Figure 8: Capturing generalizations via metafunctionality

account for the intuition that these concepts indeed function at different strata?

Lang already suggests that there may be certain genuinely 'linguistic' features that function definitionally for features at the 'conceptual' stratum:

> '... the representation of nouns like *Ofen* [oven], *Fahrzeug* [vehicle], *Boot* [boat] in the lexicon contains a specific component PURPOSE (hence, an element of our linguistic knowledge) by means of which the sort *Nutzgegenstand* [article for practical use] in the knowledge base is being accessed. This is but one example of how linguistic aspects of lexical representation can be made use of in defining ontological sorts in the knowledge base.' [Lang, 1991, p470]

Other 'genuine linguistic features' that Lang suggests for the basis of the ontological distinctions include: 'bounded object' vs. 'non-bounded object', 'concrete object' vs. 'abstract object' — both very similar to other theoretical accounts. We can now go further and *explain* the relation between the linguistic (semantic) ontology types and the conceptual ontology types as follows.

All of the reasoning that we have applied to the development of the Upper Model ontology with respect to its motivation in the lexicogrammar can be applied precisely to the relation between the Upper Model ontology and some higher stratum ontology. This follows as a consequence from the theoretical statement of the nature of realization within the stratified linguistic account. This means that we will need to find motivations for the semantic interface ontology sorts. It also means, however, that we can make use of the realization relation starting from the standpoint of the higher stratum and interpret the status of the semantic interface ontology as *generalizing* across different conceptual stratum situations; cf. Figure 5. Thus, for both the lexicogrammar with respect to the semantic interface ontology, and for the semantic interface ontology with respect to the conceptual ontology, it is likely that the *more general intra-stratal organization of the lower stratum is likely to be echoed in the overall intra-organization of the higher stratum.* This gives us the observed link between constructs that are motivateable as general semantic concepts and constructs that appear to organize the conceptual hierarchy. There is, then, no 'mixing' of the cat-

egories of different strata, there is just a resonance or echo of categories at one stratum taken up at another.

Given both this theoretical and practical binding of the contents of the different strata, is it clear why there is then a certain *tension* between strata — as [Klose and von Luck, 1991, p462] note from their experience with the LILOG ontology:

> 'The tension between linguistic and inferential demands on the modeling is alive and forces compromises on both sides.'

I have suggested that the kind of view of realization between strata found in systemic-functional linguistics, where there is both a practical *and* a theoretical 'pulling' in both directions — upwards to context and downwards to (experiential) lexicogrammar, offers an appropriate way of operating within this tension between strata. The resulting methodology then uses the tension to *help constrain* organization decisions for the construction of interface ontologies that are useful for NLP and to remove the need for genuine 'compromises' where an inappropriate category is postulated at one level because that level is insufficiently functionally differentiated from others.

It is clear, however, that we know a great deal more about possibilities for ontologies of type O_i than we do about ontologies of type O_c. Moreover, given the results of the last section, perhaps we know even less than we thought — clearly conceptual categories are now sometimes best reappropriated to a more abstract semantic type. This is a less than ideal situation — particularly given the view of stratification shown in Figure 5 and the established dialectic between strata. Because the realization relationship between strata is *bi-directional*, we should be able to use a higher-strata to constrain our accounts at a lower-strata. But the fact that we know very little about the higher-strata in this case removes one source of possible constraint.

Finally, here, however, I will draw attention to one interesting consequence for the *status* of the higher-stratum ontology when we take into account the bi-directionality of the inter-stratal relationship. Since there is no difference assumed in the theoretical status of the levels related by the interstratal relationship, one might ask how it is that the interface ontology is termed 'linguistic' and 'semantic', whereas the higher-stratum ontology is 'non-linguistic' and 'conceptual'. I believe that a far more appropriate view of the relationship is as depicted graphically in Figure 9. All strata that stand in an interstratal relationship of the kind explored and used in this paper should be seen as *semiotic* levels of greater and lesser degrees of abstraction. The conceptual ontology thus becomes more of a *contextual* ontology, with context being interpreted in the sense of a level of social situation — closely in line with, for example, [Halliday, 1978]. There is, then, the additional question of how this entire complex of inter-related levels of semiotic descriptions relates to the supporting conceptual system of human psychology. This is probably a very different kind of relation than realization — although it will probably again turn out to be a dialectic relationship rather than a one-way determination. This puts us in the position to criticise some of the conceptual sorts proposed by Lang on exactly the same grounds that he has criticised mixed ontologies. For example, alongside the above mentioned domains, all of which may be more plausibly ground in the conceptual/perceptual system, [Lang, 1991, p474] places: 'social institutions (law, administration, marriage, education)' and communicative behaviour (etiquette, conversation, group dynamics). Such a mixed set of sorts is unlikely to form a very stable or usable ontology: it is probably crucial to begin to refine further our levels of ontology, and their interactions, so that the mistakes made at the least abstract levels of ontological engineering are not just repeated again, at the next level 'up'. More detailed statements must, however, be left to future research!

6 Summary, conclusion and final words

This discussion of this paper has considered the notion of 'ontology'. Starting from the view that an ontology is an organization of the world — which has been approached by 'naive physics', 'conceptual dependencies',

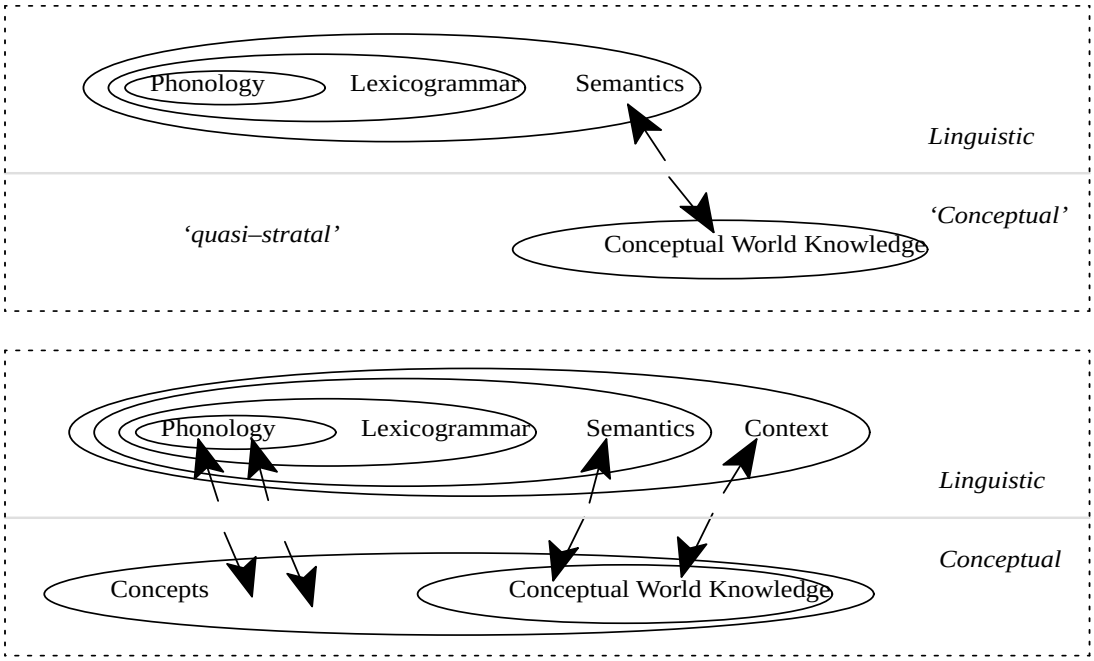

Figure 9: The relationship between semiotic descriptions and conceptual representation

'commonsense (meta)physics', and others — I drew attention to the fact that such accounts do not bring strong methodological and substantive constraints to bear on ontology construction. Also unclear is the relationship of such ontologies to language. The gap is often so large that this level is too abstract to have any direct relationship to required forms of expression. Contrariwise, this gap also leads to a weakening of the discriminative power of the constraints that can be brought to bear by linguistic patternings. Concretely, then, one cannot, for example, generate natural language directly from such levels of description without resolving, or 'fixing' an immense number of degrees of freedom that remain unaddressed (often quite rightly, if this is seen as a conceptual ontology) in the ontology itself. Much of the work that an NLP system requires to be done is, therefore, simply not taken into consideration by the abstract ontology. Such ontologies are also, because of their abstractness, difficult to populate reliably — if sizeable and potentially distributed resource construction is undertaken, as it increasingly is, then this virtually guarantees poor intercoder consistency. In short, such ontologies are of very limited value for NLP work.

These problems have been noted by some of those who have sought principles for ontology design (cf. [Skuce and Monarch, 1990]) and those who need real shareable resources (as for example in machine translation — cf. [Steiner and Reuther, 1989]). The only solution that has been found to this endeavor is to place more reliance on *language* as a source of constraint. For this reason, then, views on language and the organization of the linguistic system become crucial for ontology design that is appropriate for NLP. Moreover, only by taking views on the linguistic system that are maximally supportive of the functionalities required of ontologies can we avoid problems of lack of abstractness (i.e., being dominated by linguistic form) and problems of too much abstractness (i.e., being dominated by semantic theories of particular areas that lack connection to linguistic form). In short, ontological engineering faces the following dilemma: interface ontologies

- need to be abstract, large-scale, re-usable information classification devices,
- but they cannot be too abstract,

- or too near syntax,
- and need to be constrained from language.

The theoretical assumptions and resulting organizational decisions that I have pursued in this paper appear to offer a very practical way of preceding within this state of affairs. I have also shown that several other beneficial properties for NLP systems are derivable from the abstract organization of the linguistic system that systemic-functional theory posits.

The paper has presented for broader debate a round of discussion that begun in the context of the developing ontology of the Penman text generation system. This work, beginning with the pre-computational descriptive account, called the *Bloomington Lattice* by Halliday and Matthiessen has passed through several instantiations in computational form. Now future work will have again consider bringing together the linguistic descriptive account — reworked to a new level of detail in [Halliday and Matthiessen, to appear] — and the computational model. It is to be hoped that this approach will build on the former success of the Upper Model, simultaneously moving us in some of the directions that I raised as responses to problems with the Upper Model. Thus, I have not suggested that the Upper Model we find in Penman is the 'general solution' to ontological engineering — there are many more criticisms to be made of this ontology, again mostly concerning the extent to which it succeeds as an instantiation of the theoretical principles that underlie it. The *function* of the ontology is also more finely circumscribed than many others — but again strictly according to the underlying theory. We are not yet at a stage where an ontology can be accepted, even pragmatically for the needs of current NLP systems, as 'complete': what is more at issue is the development of appropriate methodologies for constructing ontologies, and here again constraints offered by the linguistic system are of paramount importance. The linguistic system, when viewed appropriately, gives a rich multidimensional set of constraints on adequate and appropriate designs for computational systems. The principle dimensions applied in this paper were those of *strata* and *metafunctions*. This by no means exhausts the possible input of the theory, however. For further dimensions of the theory, see [Matthiessen and Bateman, 1991]; for additional examples of using these dimensions to constrain computational system design, see [Bateman *et al.*, 1992]. I hope that the paper has suggested some of the benefits of employing such linguistic motivations, and that further attempts to apply wider sets of motivations will help us in the future.

Acknowledgments

This paper is based on the work of the natural language group at ISI, including over the years the input of Bill Mann, Ed Hovy, Christian Matthiessen, Bob Kasper, Johanna Moore, Cécile Paris, Richard Whitney, and Robert Albano. Further theoretical discussion with Erich Steiner, Jörg Schütz, and Cornelia Zelinsky-Wibbelt (IAI – Saarbrücken), Elisabeth Maier, Elke Teich, Renate Henschel and Leo Wanner (IPSI), Martin Emele and Rémi Zajac (IMS – Stuttgart) and with participants at the Technical University of Berlin International Workshop on 'Text Representation and Domain Modelling' (October 1991) and of a KIT Projekt Kolloquium (TU Berlin; February 1992) have also helped the development of the discussion significantly. The particular opinions expressed in the paper, and especially their deficiencies, remain however my responsibility.

References

[Allgayer *et al.*, 1989] Jürgen Allgayer, Karin Harbusch, Alfred Kobsa, Carola Reddig, Norbert Reithinger, and Dagmar Schmauks. Xtra: a natural-language access system to expert systems. *International Journal of Man-Machine Communication*, 1989.

[Bateman and Matthiessen, to appear] John A. Bateman and Christian M.I.M. Matthiessen. Uncovering the text base. In Hermann Bluhme and Hao Keqi, editors, *Selected Papers from the International Conference on Research in Text and Language, Xi'an Jiaotong University, Xi'an, P.R. China, 29-31 March 1989*. Xi'an Jiaotong University Press, Xi'an, People's Republic of China, to appear.

[Bateman and Paris, 1] John A. Bateman and Cécile L. Paris. Phrasing a text in terms the user can understand. In *Proceedings of the Eleventh International Joint Conference on*

Artificial Intelligence, Detroit, Michigan, 1. IJCAI-89.

[Bateman *et al.*, 1990] John A. Bateman, Robert T. Kasper, Johanna D. Moore, and Richard A. Whitney. A general organization of knowledge for natural language processing: the PENMAN upper model. Technical report, USC/Information Sciences Institute, Marina del Rey, California, 1990.

[Bateman *et al.*, 1991] John A. Bateman, Christian M.I.M. Matthiessen, Keizo Nanri, and Licheng Zeng. Multilingual text generation: an architecture based on functional typology. In *International Conference on Current Issues in Computational Linguistics*, Penang, Malaysia, 1991. Also available as technical report of the department of Linguistics, University of Sydney.

[Bateman *et al.*, 1992] John Bateman, Elisabeth Maier, Christian Matthiessen, and Cecile Paris. Generation systems design: Issues of modularity. Technical report, GMD, Integrated Publication and Information Systems Institute, Darmstadt, Germany, 1992. forthcoming.

[Bateman, 1989] John A. Bateman. Upper modelling for machine translation: a level of abstraction for preserving meaning. Technical Report EUROTRA-D Working Papers, No. 12, Institut für Angewandte Informationsforschung, Saarbrücken, West Germany, 1989.

[Bateman, 1990a] John A. Bateman. Finding translation equivalents: an application of grammatical metaphor. In *13th. International Conference on Computational Linguistics (COLING-90)*, volume II, pages 13–18, Helsinki, Finland, 1990.

[Bateman, 1990b] John A. Bateman. Upper modeling: organizing knowledge for natural language processing. In *5th. International Workshop on Natural Language Generation, 3-6 June 1990*, Pittsburgh, PA., 1990. Organized by Kathleen R. McKeown (Columbia University), Johanna D. Moore (University of Pittsburgh) and Sergei Nirenburg (Carnegie Mellon University).

[Bateman, 1991] John A. Bateman. Uncovering textual meanings: a case study involving systemic-functional resources for the generation of Japanese texts. In Cécile L. Paris, William R. Swartout, and William C. Mann, editors, *Natural language generation in artificial intelligence and computational linguistics*. Kluwer Academic Publishers, July 1991. Presented at the Fourth International Workshop on Natural Language Generation. Santa Catalina Island, California, July, 1988.

[Bierwisch and Lang, 1989] Manfred Bierwisch and Ewald Lang, editors. *Dimensional adjectives. Grammatical structure and conceptual interpretation*. Springer-Verlag, Berlin, 1989.

[Bierwisch, 1982] M. Bierwisch. Semantische und konzeptuelle Repräsentation lexikalischer Einheiten. In R. Ružička and W. Motsch, editors, *Untersuchungen zur Semantik*, pages 61 – 99. Akademie-Verlag, Berlin, 1982.

[Bosch, 1991] Peter Bosch. The Bermuda triangle: natural language semantics between linguistics, knowledge representation, and knowledge processing. In O. Herzog and C.-R. Rollinger, editors, *Text understanding in* LILOG*: integrating computational linguistics and artificial intelligence, Final report on the IBM Germany* LILOG*-Project*, pages 243 – 258. Springer-Verlag, Berlin, 1991. Lecture notes in artificial intelligence, 546.

[Brachman and Schmolze, 1985] Ronald J. Brachman and J. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2), 1985.

[Calder *et al.*, 1989] Jo Calder, Mike Reape, and Henk Zeevat. An algorithm for generation in unification categorial grammar. In *Proceedings of the 4th Conference of the European Chapter of the Association of Computational Linguistics*, Manchester, England, 1989. Association for Computational Linguistics.

[Calzolari, 1991] Nicolleta Calzolari. Acquiring and representing semantic information in a Lexical Knowledge Base. In James Pustejovsky and Sabine Bergler, editors, *Proceedings of 1991 ACL Workshop on Lexical Semantics and Knowledge Representation*, pages 188–197, 1991.

[Carbonell and Tomita, 1987] Jaime G. Carbonell and Masaru Tomita. Knowledge-based machine translation, the CMU approach. In Sergei Nirenburg, editor, *Theoretical and Methodological Issues in Machine Translation*, pages 68–89. Cambridge University Press, Cambridge, 1987.

[Chen and Cha, 1988] Keh-Jiann Chen and Chuan-Shu Cha. The design of a conceptual structure and its relation to the parsing of chinese sentences. In *Proceedings of the 1988 International Conference on Computer Processing of Chinese and Oriental Languages*, Toronto, Canada, August 29 - September 1 1988.

[Chomsky, 1980] Noam Chomsky. On binding. *Linguistic Inquiry*, 11(1):1–46, 1980.

[Dahlgren *et al.*, 1989] Kathleen Dahlgren, Joyce McDowell, and Edward P. Stabler. Knowledge representation for commonsense reasoning with text. *Computational Linguistics*, 15(3):149–170, 1989.

[Dorr, 1987] Bonnie J. Dorr. Unitran: a principle-based approach to machine translation. Technical Report MIT AI Technical Report 1000, Massachussetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, Massachussetts, 1987.

[Dorr, 1990] Bonnie J. Dorr. *Lexical conceptual structure and machine translation*. PhD thesis, Massachussetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, Massachussetts, 1990.

[Dorr, 1991] Bonnie J. Dorr. A two-level knowledge representation for machine translation: lexical semantics and tense/aspect. In James Pustejovsky and Sabine Bergler, editors, *Proceedings of the 1991 ACL Workshop on Lexical Semantics and Knowledge Representation*, pages 250 – 263, Berkeley, CA, June 1991.

[Emele *et al.*, 1990] Martin Emele, Ulrich Heid, Walter Kehl, Stefan Momma, and Rémi Zajac. Organizing linguistic knowledge for multilingual generation. Technical report, Project Polygloss, University of Stuttgart, West Germany, 1990. Paper submitted to COLING-90.

[Emele, 1989] Martin C. Emele. A typed-feature structure unification-based approach to generation. In *Proceedings of the WGNLC of the IECE*, Oita University, Japan, 1989.

[Fawcett, 1987] Robin P. Fawcett. The semantics of clause and verb for relational processes. In Robin P. Fawcett and David J. Young, editors, *New Developments in Systemic Linguistics: Volume 1*. Frances Pinter, London, 1987.

[Fillmore, 1968] Charles J. Fillmore. The case for case. In Emons Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*. Holt, Rinehart and Wilson, New York, 1968.

[Gust, 1991] Helmar Gust. Representing word meanings. In O. Herzog and C.-R. Rollinger, editors, *Text understanding in* LILOG*: integrating computational linguistics and artificial intelligence, Final report on the IBM Germany* LILOG*-Project*, pages 127 – 142. Springer-Verlag, Berlin, 1991. Lecture notes in artificial intelligence, 546.

[Hale and Keyser, 1986] K. Hale and J. Keyser. Some transitivity alternations in english. Technical Report Lexicon Project Working Papers 7, Massachussetts Institute of Technology, Cambridge, Massachussetts, 1986.

[Halliday and Matthiessen, to appear] Michael A.K. Halliday and Christian M.I.M. Matthiessen. *Construing experience through meaning: a language-based approach to cognition*. de Gruyter, Berlin, to appear.

[Halliday, 1961] Michael A.K. Halliday. Categories of the theory of grammar. *Word*, 17:241–292, 1961. Reprinted in abbreviated form in Halliday (1976) edited by Gunther R. Kress, pp 52-72.

[Halliday, 1978] Michael A.K. Halliday. *Language as social semiotic*. Edward Arnold, London, 1978.

[Halliday, 1982] Michael A.K. Halliday. How is a text like a clause? In Sture Allén, editor, *Text Processing*. Almqvist and Wiksell, Stockholm, 1982.

[Halliday, 1985] Michael A.K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London, 1985.

[Hayes, 1979] Patrick J. Hayes. The naive physics manifesto. In Donald Michie, editor, *Expert systems in the microelectronic age*. Edinburgh University Press, Edinburgh, Scotland, 1979.

[Hayes, 1985] Patrick J. Hayes. Native physics i: ontology for liquids. In Jerry R. Hobbs and R.C. Moore, editors, *Formal theories of the commonsense world*, pages 71 – 108. Ablex Publishing Corporation, New Jersey, 1985.

[Heid *et al.*, 1988] Ulrich Heid, Dietmar Rösner, and Birgit Roth. Generating german from semantic relations: Semantic relations as an input to the SEMSYN generator. In Erich H. Steiner, Paul Schmidt, and Cornelia Zelinksy-Wibbelt, editors, *From Syntax to Semantics: experiences from Machine Translation*. Frances Pinter, London, 1988.

[Herskovits, 1986] A. Herskovits. *Language and Spatial Cognition: an interdisciplinary study of the prepositions in English*. Cambridge, 1986.

[Herweg, 1991] Michael Herweg. Aspectual requirements of temporal connectives: evidence for a two-level approach to semantics. In James Pustejovsky and Sabine Bergler, editors, *Proceedings of the 1991 ACL Workshop on Lexical Semantics and Knowledge Representation*, pages 152 – 164, Berkeley, CA, June 1991.

[Herzog and Rollinger, 1991] Otthein Herzog and Claus-Rainer Rollinger, editors. *Text understanding in* LILOG*: integrating computational linguistics and artificial intelligence, Final report on the IBM Germany* LILOG*-Project*. Springer-Verlag, Berlin, 1991. Lecture notes in artificial intelligence, 546.

[Hinrichs *et al.*, 1987] E.W. Hinrichs, D.M. Ayuso, and R. Scha. *The syntax and semantics of the* JANUS *semantic interpretation language*, pages 27–31. BBN Laboratories, Report No. 6552, 1987.

[Hobbs and Moore, 1985] Jerry R. Hobbs and R.C. Moore, editors. *Formal theories of the commonsense world*. Ablex Publishing Corporation, New Jersey, 1985.

[Hobbs *et al.*, 1987] Jerry R. Hobbs, William Croft, Todd Davies, Douglas Edwards, and Kenneth Laws. Commonsense metaphysics and lexical semantics. *Computational Linguistics*, 13(3-4):241 – 250, July - December 1987.

[Hobbs, 1985] Jerry R. Hobbs. On the coherence and structure of discourse. Technical Report CLSI-85-37, Center for the Study of Language and Information, Stanford, CA, October 1985.

[Horacek, 1989] Helmut Horacek. Towards principles of ontology. In D. Metzing, editor, *Proceedings of the German Workshop on Artificial Intelligence: GWAI89*, pages 323 – 330. Springer-Verlag, Berlin, Heidelberg, New York, 1989.

[Hovy, 1988] Eduard H. Hovy. *Generating natural language under pragmatic constraints*. Lawrence Erlbaum, Hillsdale, New Jersey, 1988.

[Jackendoff, 1977] Ray Jackendoff. $\overline{X}$ *Syntax: a study of phrase structure*. The M.I.T. Press, Cambridge, MA, 1977.

[Jackendoff, 1983] Ray Jackendoff. *Semantics and Cognition*. The M.I.T. Press, Cambridge, MA, 1983.

[Jackendoff, 1987] Ray Jackendoff. The status of thematic relations in linguistic theory. *Linguistic Inquiry*, 18(3):369–411, 1987.

[Jackendoff, 1990] Ray Jackendoff. *Semantic Structures*. The M.I.T. Press, Cambridge, MA, 1990.

[Klose and von Luck, 1991] Gudrun Klose and Kai von Luck. The background knowledge of the LILOG system. In O. Herzog and C.-R. Rollinger, editors, *Text understanding in* LILOG*: integrating computational linguistics and artificial intelligence, Final report on the IBM Germany* LILOG*-Project*, pages 455 – 463. Springer-Verlag, Berlin, 1991. Lecture notes in artificial intelligence, 546.

[Klose *et al.*, 1991] Gudrun Klose, Ewald Lang, and Thomas Pirlein. Die Ontologie und Axiomatik der Wissensbasis von LEU/2: Erfahrungen – Probleme – Ausblicke. Technical Report IWBS Report 171, IBM Deutschland, Stuttgart, 1991.

[Krifka, 1989] Manfred Krifka. Nominalreferenz, Zeitkonstitution, Aspekt, Aktionsart: Eine semantische Erklärung ihrer Interpretation. In W. Abraham and T. Janssen, editors, *Tempus—Aspekt—Modus*. Niemeyer, Tübingen, 1989.

[Lang, 1991] Ewald Lang. The LILOG ontology from a linguistic point of view. In O. Herzog and C.-R. Rollinger, editors, *Text understanding in* LILOG*: integrating computational linguistics and artificial intelligence, Final report on the IBM Germany* LILOG*-Project*, pages 464 – 481. Springer-Verlag, Berlin, 1991. Lecture notes in artificial intelligence, 546.

[Langacker, 1987] Ronald W. Langacker. *Foundations in Cognitive Grammar*. Stanford University Press, Stanford, California, 1987.

[Lenat and Guha, 1988] Doug Lenat and R.V. Guha. The world according to CYC. Technical Report MCC Technical Report ACA-AI-300-88, Microelectronics and Computer Technology Corporation, Austin, Texas, September 1988.

[Levin, 1987] Lori Levin. Towards a linking theory of relation changing rules in lfg. Technical Report Report No. CSLI-87-115, Center for the Study of Language and Information, Stanford, California, 1987.

[Mann and Matthiessen, 1985] William C. Mann and Christian M.I.M. Matthiessen. Demonstration of the Nigel text generation computer program. In James D. Benson and William S. Greaves, editors, *Systemic Perspectives on Discourse, Volume 1*. Ablex, Norwood, New Jersey, 1985.

[Mann *et al.*, 1985] William C. Mann, Yigal Arens, Christian M.I.M. Matthiessen, Shari Naberschnig, and Norman K. Sondheimer. Janus abstraction structure — draft 2. Technical report, USC/Information Sciences Institute, Marina del Rey, California, October 1985. (Circulated in draft form only.).

[Mann, 1983] William C. Mann. The anatomy of a systemic choice. *Discourse Processes*, 1983. Also available as USC/Information Sciences Institute, Research Report ISI/RR-82-104, 1982.

[Mann, 1985] William C. Mann. Janus abstraction structure – draft 1, 1985. An informal project technical memo of the Janus project at ISI.

[Matsukawa and Yokota, 1991] Tomoyoshi Matsukawa and Eiji Yokota. Development of the concept dictionary – implementation of lexical knowledge. In James Pustejovsky and Sabine Bergler, editors, *Proceedings of the ACL Workshop on Lexical Semantics and Knowledge*

Representation, pages 206 –223, Berkeley, CA, June 1991.

[Matthiessen and Bateman, 1991] Christian M.I.M. Matthiessen and John A. Bateman. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Frances Pinter Publishers and St. Martin's Press, London and New York, 1991.

[Matthiessen and Halliday, forthcoming] Christian M.I.M. Matthiessen and Michael A.K. Halliday. Systemic Functional Grammar. In Fred C.C. Peng and J. Ney, editors, *Current Approaches to Syntax*. Benjamins and Whurr, Amsterdam and London, forthcoming.

[Matthiessen, 1990] Christian M.I.M. Matthiessen. Lexicogrammatical cartography: English systems. Technical report, University of Sydney, Linguistics Department, 1990. Ongoing expanding draft.

[McKeown and Paris, 1987] Kathleen R. McKeown and Cécile L. Paris. Functional unification grammar revisited. In *Proceedings of the 25th Annual Meeting of the ACL*, Palo Alto, California, 1987. Association of Computational Linguistics.

[Meteer *et al.*, 1987] Marie W. Meteer, David D. McDonald, S.D. Anderson, D. Forster, L.S. Gay, A.K. Huettner, and P. Sibun. MUMBLE-86: Design and implementation. Technical Report 87-87, COINS, University of Massachusetts, 1987.

[Meteer, 1988] Marie W. Meteer. Defining a vocabulary for text planning, August 1988. Presented at the AAAI-88 Workshop on Text Planning and Realization, organized by Eduard H. Hovy, Doug Appelt, David McDonald and Sheryl Young.

[Meteer, 1989] Marie W. Meteer. The SPOKESMAN natural language generation system. Technical Report BBN Report No. 7090, BBN Systems and Technologies Corporation, Cambridge, MA, 1989.

[Moens and Steedman, 1988] M. Moens and M. Steedman. Temporal ontology and temporal reference. *Computuational Linguistics*, 14(2), 1988.

[Moens *et al.*, 1989] Marc Moens, Jo Calder, Ewan Klein, Mark Reape, and Henk Zeevat. Expressing generalizations in unification-based formalisms. In *Proceedings of the 4th. Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Manchester, England, 1989.

[Momma and Dörre, 1987] Stefan Momma and Jochen Dörre. Generation from f-structures. In Ewan Klein and Johann Van Bentham, editors, *Categories, Polymorphism and Unification*. Cognitive Science Centre, University of Edinburgh, Edinburgh, Scotland, 1987.

[Moore and Arens, 1985] Johanna D. Moore and Yigal Arens. A hierarchy for entities, 1985. USC/Information Sciences Institute, Internal Draft.

[Moore, 1989] Johanna D. Moore. *A reactive approach to explanation in expert and advice-giving systems*. PhD thesis, University of California, Los Angeles, 1989.

[Nebel *et al.*, 1991] Bernhard Nebel, Christof Peltason, and Kai von Luck. Proceedings of international workshop on terminological logics. Technical Report DFKI-D-91-13, DFKI, Saarbrücken, May 1991.

[Nerbonne, 1992] John Nerbonne. Representing grammar, meaning and knowledge, May 1992. (Papers from KIT-FAST Workshop, Technical University Berlin, October 9th - 11th 1991).

[Nirenburg and Levin, 1991] Sergei Nirenburg and Levin. Syntax-driven and ontology-driven lexical semantics. In James Pustejovsky and Sabine Bergler, editors, *Proceedings of the ACL Workshop on Lexical Semantics and Knowledge Representation*, Berkeley, CA, June 1991.

[Nirenburg and Raskin, 1987] Sergei Nirenburg and Victor Raskin. The subworld concept lexicon and the lexicon management system. *Computational Linguistics*, 13(3-4), 1987.

[Nirenburg *et al.*, 1987] Sergei Nirenburg, V. Raskin, and A. Tucker. The structure of interlingua in TRANSLATOR. In Sergei Nirenburg, editor, *Machine Translation: Theoretical and Methodological Issues*. Cambridge University Press, Cambridge, 1987.

[Onyshvekych and Nirenburg, 1991] Boyan A. Onyshvekych and Sergei Nirenburg. Lexicon, ontology and text meaning. In James Pustejovsky and Sabine Bergler, editors, *Proceedings of the ACL Workshop on Lexical Semantics and Knowledge Representation*, Berkeley, CA, June 1991.

[Penman Project, 1989] Penman Project. PENMAN documentation: the Primer, the User Guide, the Reference Manual, and the Nigel manual. Technical report, USC/Information Sciences Institute, Marina del Rey, California, 1989.

[Pirlein, 1991] Thomas Pirlein. Konstruktion und Evaluation von Wissensbasen in textverstehenden Systemen. In Th. Christaller, editor, *GWAI-91: 15. Fachtagung für Künstliche*

Intelligenz, pages 147 –156. Springer-Verlag, Berlin, 1991.

[Pollard and Sag, 1987] Carl Pollard and Ivan A. Sag. *Information-based syntax and semantics: volume 1*. Chicago University Press, Chicago, 1987. Center for the Study of Language and Information; Lecture Notes Number 13.

[Pustejovsky, 1988] James Pustejovsky. Event semantic structure. Technical report, Brandeis University, Waltham, MA., 1988.

[Pustejovsky, 1991] James Pustejovsky. Towards a generative lexicon. *Computational Linguistics*, 17(4), 1991.

[Reinhardt and Whipple, 1988] T. Reinhardt and C. Whipple. Summary of conclusions from the longman's taxonomy experiment. In B. Goodman, editor, *Annual Report*. BBN Systems and Technologies Corporation, Cambridge, MA, 1988.

[Rohrer, 1986] Christian Rohrer. Linguistic bases for machine translation. In *Proceedings of* COLING *86*, pages 353–355, 1986. 11th. International Conference on Computational Linguistics; Bonn, August.

[Rösner, 1988] Dietmar Rösner. The generation system of the SEMSYN project: towards a task-independent generator for German. In Michael Zock and Gérard Sabah, editors, *Advances in Natural Language Generation: an interdisciplinary perspective; volume 2*. Frances Pinter, London, 1988.

[Sag and Pollard, 1991] Ivan A. Sag and Carl Pollard. An integrated theory of complement control. *Language*, 67(1):63 – 113, 1991.

[Schank and Abelson, 1977] Roger C. Schank and R. P. Abelson. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

[Schriefers, 1990] Heribert Schriefers. Lexical and conceptual factors in the naming of relations. *Cognitive Psychology*, 22:111 – 142, 1990.

[Simmons, 1991] G. Simmons. Empirical methods for ontological engineering. In Gudrun Klose, Ewald Lang, and Thomas Pirlein, editors, *Die Ontologie und Axiomatik der Wissensbasis von LEU/2: Erfahrungen – Probleme – Ausblicke*, number IWBS Report 171, pages VI. 1 – 38. 1991.

[Skuce and Monarch, 1990] Doug Skuce and Ira Monarch. Ontological issues in knowledge base design: some problems and suggestions. In *Proceedings of the Banff Workshop on Knowledge Acquisition*, Banff, Canada, 1990.

[Smolka and Aït-Kaci, 1989] Gert Smolka and Hassan Aït-Kaci. Inheritance hierarchies: semantics and unification. *Journal of symbolic computation*, 7:343 – 370, 1989.

[Smolka, 1989] Gert Smolka. A feature logic with subsorts. Technical Report LILOG Report, 33, IWBS, IBM Deutschland, Postfach 80 08 80, Stuttgart, 1989. (To appear in the Proceedings of the Workshop on Unification Formalisms – Syntax, Semantics, and Implementation, Titisee, The MIT Press, 1990.).

[Steiner and Reuther, 1989] Erich H. Steiner and Ursula Reuther. Semantic relations. EUROTRA reference manual, Version 6, 1989. (Commission of the European Community).

[Steiner *et al.*, 1987] Erich H. Steiner, Ursula Eckert, Birgit Weck, and Jutta Winter. The development of the EUROTRA-D system of semantic relations. Technical Report Eurotra-D Working Papers, No. 2, Institut der angewandten Informationsforschung, Universität des Saarlandes, Saarbrücken, West Germany, 1987.

[Steiner, 1987] Erich H. Steiner. Semantic relations in lfg and EUROTRA-D — a comparison. Technical Report EUROTRA-D Working Papers No. 5, Institut für Angewandte Informationsforschung, Saarbrücken, West Germany, 1987.

[Talmy, 1987] Leonard Talmy. The relation of grammar to cognition. In B. Rudzka-Ostyn, editor, *Topics in Cognitive Linguistics*. John Benjamins, 1987.

[Vendler, 1967] Z. Vendler. *Linguistics in Philosophy*. Cornell University Press, Ithaca, 1967.

[Weischedel, 1989] Ralph M. Weischedel. A hybrid approach to representation in the JANUS natural language processor. In *27th Annual Meeting of the Association for Computational Linguistics, 26-29 June 1989*, pages 193–202, Vancouver, British Columbia, 1989. The Association for Computational Linguistics.

[Wierzbicka, 1988] Anna Wierzbicka. *The semantics of grammar*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1988. Studies in Language Companion Series 18.

[Zajac, 1989] Rémi Zajac. A transfer model using a typed feature structure rewriting system with inheritance. In *27th Annual Meeting of the Association for Computational Linguistics, 26-29 June 1989*, pages 193–202, Vancouver, British Columbia, 1989. The Association for Computational Linguistics.

[Zelinsky-Wibbelt, 1987] Cornelia Zelinsky-Wibbelt. Semantische Merkmale für die automatische Disambiguierung: ihre Generierung

und ihre Verwendung. Technical Report EUROTRA-D Working Papers No. 4, Institut für Angewandte Informationsforschung, Eurotra-D, Saarbrücken, West Germany, 1987.

[Zelinsky-Wibbelt, 1988]
Cornelia Zelinsky-Wibbelt. The semantic representation of sentences by means of semantic features. In Erich Steiner, Paul Schmidt, and Cornelia Zelinksy-Wibbelt, editors, *From Syntax to Semantics: insights from Machine Translation*. Frances Pinter, London, 1988.